

Robert Meersman
Zahir Tari
Pilar Herrero et al. (Eds.)

LNCS 4277

On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops

**OTM Confederated International Workshops and Posters
AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET,
OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium,
RDDS, SWWS, and SeBGIS 2006
Montpellier, France, October/November 2006, Proceedings, Part I**

1 Part I

 **Springer**

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

University of Dortmund, Germany

Madhu Sudan

Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Robert Meersman Zahir Tari
Pilar Herrero et al. (Eds.)

On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops

OTM Confederated International Workshops and Posters
AWeSOMe, CAMS, COMINF, IS, KSinBIT, MIOS-CIAO, MONET,
OnToContent, ORM, PerSys, OTM Academy Doctoral Consortium,
RDDS, SWWS, and SeBGIS 2006
Montpellier, France, October 29 – November 3, 2006
Proceedings, Part I

Volume Editors

Robert Meersman
Vrije Universiteit Brussel (VUB), STARLab
Bldg G/10, Pleinlaan 2, 1050 Brussels, Belgium
E-mail: meersman@vub.ac.be

Zahir Tari
RMIT University, School of Computer Science and Information Technology
Bld 10.10, 376-392 Swanston Street, VIC 3001, Melbourne, Australia
E-mail: zahirt@cs.rmit.edu.au

Pilar Herrero
Universidad Politécnica de Madrid, Facultad de Informática
Campus de Montegancedo S/N, 28660 Boadilla del Monte, Madrid, Spain
E-mail: pherrero@fi.upm.es

Library of Congress Control Number: 2006935257

CR Subject Classification (1998): H.2, H.3, H.4, C.2, H.5, I.2, D.2, K.4

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-540-48269-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-48269-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11915034 06/3142 5 4 3 2 1 0

Volume Editors

Robert Meersman
Zahir Tari
Pilar Herrero

AWeSOMe

Daniel Grosu
Pilar Herrero
Gonzalo Medez
Marta Sabou

CAMS

Annika Hinze
George Buchanan

COMINF

Aldo de Moor
Michael Gurstein

IS

Mario Freire
Simao Melo de Sousa
Vitor Santos

KSinBIT

Maja Hadzic
Bart De Moor
Yves Moreau
Arek Kasprzyk

MIOS-CIAO

Antonia Albani
Jan L.G. Dietz

MONET

Fernando Ferri
Maurizio Rafanelli
Arianna D'Ulizia

OnToContent

Mustafa Jarrar
Claude Ostyn
Werner Ceusters
Andreas Persidis

ORM

Terry Halpin
Robert Meersman

PerSys

Skevos Evripidou
Roy Campbell

OTM Academy Doctoral Consortium

Antonia Albani
Gabor Nagypal
Johannes Maria Zaha

RDDS

Eiko Yoneki
Pascal Felber

SWWS

Katia Sycara
Elizabeth Chang
Ernesto Damiani
Mustafa Jarrar
Tharam Dillon

OTM 2006 General Co-chairs' Message

Dear OnTheMove Participant or Reader of these Proceedings,

The General Chairs of OnTheMove 2006, Montpellier, France, are happy to observe that the conference series that was started in Irvine, California in 2002 and subsequently held in Catania, Sicily in 2003 and in Cyprus in 2004 and 2005 clearly continues to attract a growing representative selection of today's world-wide research on the scientific concepts underlying distributed, heterogeneous and autonomous yet meaningfully collaborative computing, with the Internet and the WWW as its prime epitomes.

Indeed, as such large, complex and networked intelligent information systems become the focus and norm for computing, it is clear that there is an acute and increasing need to address and discuss in an integrated forum the implied software and system issues as well as methodological, theoretical and application issues. As we all know, e-mail, the Internet, and even video conferences are not sufficient for effective and efficient scientific exchange. This is why the OnTheMove (OTM) Federated Conferences series has been created to cover the increasingly wide yet closely connected range of fundamental technologies such as data and Web semantics, distributed objects, Web services, databases, information systems, workflow, cooperation, ubiquity, interoperability, mobility, grid and high-performance. OnTheMove aspires to be a primary scientific meeting place where all aspects of the development of Internet- and Intranet-based systems in organizations and for e-business are discussed in a scientifically motivated way. This fifth 2006 edition of the OTM Federated Conferences event therefore again provided an opportunity for researchers and practitioners to understand and publish these developments within their individual as well as within their broader contexts.

The backbone of OTM was originally formed by the co-location of three related, complementary and successful main conference series: DOA (Distributed Objects and Applications, since 1999), covering the relevant infrastructure-enabling technologies, ODBASE (Ontologies, DataBases and Applications of SEmantics, since 2002) covering Web semantics, XML databases and ontologies, CoopIS (Cooperative Information Systems, since 1993) covering the application of these technologies in an enterprise context through, for example, workflow systems and knowledge management. For the 2006 edition, these were strengthened by a fourth conference, GADA (Grid computing, high-performAnce and Distributed Applications, a successful workshop at OTM since 2004), covering the large-scale integration of heterogeneous computing systems and data resources with the aim of providing a global computing space. Each of these four conferences encourages researchers to treat their respective topics within a framework that incorporates jointly (a) theory, (b) conceptual design and development, and (c) applications, in particular case studies and industrial solutions.

Following and expanding the model created in 2003, we again solicited and selected quality workshop proposals to complement the more “archival” nature of the main conferences with research results in a number of selected and more “avant garde” areas related to the general topic of distributed computing. For instance, the so-called Semantic Web has given rise to several novel research areas combining linguistics, information systems technology, and artificial intelligence, such as the modeling of (legal) regulatory systems and the ubiquitous nature of their usage. We were glad to see that several earlier successful workshops (notably WOSE, MIOS-INTEROP, AweSOMe, CAMS, SWWS, SeBGIS, ORM) re-appeared in 2006 with a second, third or sometimes fourth edition, and that not less than seven new workshops could be hosted and successfully organized by their respective proposers: IS (International Workshop on Information Security), COMINF (International Workshop on Community Informatics), KSinBIT (International Workshop on Knowledge Systems in Bioinformatics), MONET (International Workshop on MOBILE and NETworking Technologies for social applications), OnToContent (Ontology content and evaluation in Enterprise), PerSys (International Workshop on Pervasive Systems), and RDDS (International Workshop on Reliability in Decentralized Distributed Systems). We know that as before, their audiences will mutually productively mingle with those of the main conferences, as is already visible from the overlap in authors! The OTM organizers are especially grateful for the leadership and competence of Pilar Herrero in managing this complex process into a success for the second year in a row.

A special mention for 2006 is again due for the third and enlarged edition of the highly attractive OnTheMove Academy (formerly called Doctoral Consortium Workshop). Its 2006 Chairs, Antonia Albani, Gábor Nagypál and Johannes Maria Zaha, three young and active researchers, further refined the original set-up and interactive formula to bring PhD students together: they call them to submit their research proposals for selection; the resulting submissions and their approaches are presented by the students in front of a wider audience at the conference, where they are then independently and extensively analyzed and discussed in public by a panel of senior professors. This year these were Johann Eder, Maria Orłowska, and of course Jan Dietz, the Dean of the OnTheMove Academy, who provided guidance, support and help for the team. The successful students are also awarded free access to all other parts of the OTM program, and only pay a minimal fee for the Doctoral Symposium itself (in fact their attendance is largely sponsored by the other participants!). The OTM organizers expect to further expand the OnTheMove Academy in future editions of the conferences and so draw an audience of young researchers into the OTM forum.

All four main conferences and the associated workshops share the distributed aspects of modern computing systems, and the resulting application-pull created by the Internet and the so-called Semantic Web. For DOA 2006, the primary emphasis was on the distributed object infrastructure; for ODBASE 2006, it became the knowledge bases and methods required for enabling the use of formal semantics; for CoopIS 2006, the topic was the interaction of such

technologies and methods with management issues, such as occur in networked organizations, and for GADA 2006, the topic was the scalable integration of heterogeneous computing systems and data resources with the aim of providing a global computing space. These subject areas naturally overlap and many submissions in fact also treat an envisaged mutual impact among them. As for the earlier editions, the organizers wanted to stimulate this cross-pollination by a shared program of famous keynote speakers: this year we were proud to announce Roberto Cencioni (European Commission), Alois Ferscha (Johannes Kepler Universität), Daniel S. Katz (Louisiana State University and Jet Propulsion Laboratory), Frank Leymann (University of Stuttgart), and Marie-Christine Rousset (University of Grenoble)! We also encouraged multiple event attendance by providing all authors, also those of workshop papers, with free access or discounts to one other conference or workshop of their choice.

We received a total of 361 submissions for the four main conferences and an impressive 493 (compared to the 268 in 2005 and 170 in 2004!) submissions for the workshops. Not only may we indeed again claim success in attracting an increasingly representative volume of scientific papers, but such a harvest of course allows the Program Committees to compose a higher quality cross-section of current research in the areas covered by OTM. In fact, in spite of the larger number of submissions, the Program Chairs of each of the three main conferences decided to accept only approximately the same number of papers for presentation and publication as in 2003, 2004 and 2005 (i.e., average one paper out of four submitted, not counting posters). For the workshops, the acceptance rate varies but was much stricter than before, about one in two to three, to less than one quarter for the IS (Information Security) international workshop. Also for this reason, we separated the proceedings into two books with their own titles, with the main proceedings in two volumes, and we are grateful to Springer for their suggestions and collaboration in producing these books and CDROMs. The reviewing process by the respective Program Committees as usual was performed very professionally and each paper in the main conferences was reviewed by at least three referees, with arbitrated e-mail discussions in the case of strongly diverging evaluations. It may be worthwhile to emphasize that it is an explicit OnTheMove policy that all conference Program Committees and Chairs make their selections completely autonomously from the OTM organization itself. Continuing a costly but nice tradition, the OnTheMove Federated Event organizers decided again to make all proceedings available to all participants of conferences and workshops, independently of one's registration to a specific conference or workshop. Each participant also received a CDROM with the full combined proceedings (conferences + workshops).

The General Chairs are once more especially grateful to all the many people directly or indirectly involved in the setup of these federated conferences who contributed to making it a success. Few people realize what a large number of people have to be involved, and what a huge amount of work, and sometimes risk, the organization of an event like OTM entails. Apart from the persons in the roles mentioned above, we therefore in particular wish to thank our 12 main conference

PC Co-chairs (GADA 2006: Pilar Herrero, María S. Pérez, Domenico Talia, Albert Zomaya; DOA 2006: Judith Bishop, Kurt Geihs; ODBASE 2006: Maurizio Lenzerini, Erich Neuhold, V.S. Subrahmanian; CoopIS 2006: Mike Papazoglou, Louiqa Raschid, Rainer Ruggaber) and our 36 workshop PC Co-chairs (Antonia Albani, George Buchanan, Roy Campbell, Werner Ceusters, Elizabeth Chang, Ernesto Damiani, Jan L.G. Dietz, Pascal Felber, Fernando Ferri, Mario Freire, Daniel Grosu, Michael Gurstein, Maja Hadzic, Pilar Herrero, Terry Halpin, Annika Hinze, Skevos Evripidou, Mustafa Jarrar, Arek Kasprzyk, Gonzalo Méndez, Aldo de Moor, Bart De Moor, Yves Moreau, Claude Ostyn, Andreas Persidis, Maurizio Rafanelli, Marta Sabou, Vitor Santos, Simao Melo de Sousa, Katia Sycara, Arianna D’Ulizia, Eiko Yoneki, Esteban Zimányi).

All, together with their many PC members, did a superb and professional job in selecting the best papers from the large harvest of submissions.

We also heartily thank Zohra Bellahsene of LIRMM in Montpellier for the considerable efforts in arranging the venue at their campus and coordinating the substantial and varied local facilities needed for a multi-conference event such as ours. And we must all also be grateful to Mohand-Said Hacid of the University of Lyon for researching and securing the sponsoring arrangements, to Gonzalo Méndez, our excellent Publicity Chair, to our extremely competent and experienced Conference Secretariat and technical support staff Daniel Meersman, Ana-Cecilia Martinez Barbosa, and Jan Demey, and last but not least to our hyperactive Publications Chair and loyal collaborator of many years, Kwong Yuen Lai, this year bravely assisted by Peter Dimopoulos.

The General Chairs gratefully acknowledge the academic freedom, logistic support and facilities they enjoy from their respective institutions, Vrije Universiteit Brussel (VUB) and RMIT University, Melbourne, without which such an enterprise would not be feasible.

We do hope that the results of this federated scientific enterprise contribute to your research and your place in the scientific network... We look forward to seeing you again at next year’s edition!

August 2006

Robert Meersman, Vrije Universiteit Brussel, Belgium
 Zahir Tari, RMIT University, Australia
 (General Co-chairs, OnTheMove 2006)

Organization Committee

The OTM (On The Move) Federated Workshops aim at complementing the more “archival” nature of the OTM Federated Conferences with research results in a number of selected and more “avant garde” areas related to the general topic of distributed computing. In 2006, only 14 workshops were chosen after a rigorous selection process by Pilar Herrero. The 2006 selected international conferences were: AWeSOMe (International Workshop on Agents, Web Services and Ontologies Merging), CAMS (International Workshop on Context-Aware Mobile Systems), COMINF (International Workshop on Community Informatics), IS (International Workshop on Information Security), KSinBIT (International Workshop on Knowledge Systems in Bioinformatics), MIOS+CIAO (International Workshop on Inter-organizational Systems), MONET (International Workshop on MOBILE and NETWORKING Technologies for social applications), OnToContent (International Workshop on Ontology content and evaluation in Enterprise), ORM (International Workshop on Object-Role Modeling), PerSys (International Workshop on Pervasive Systems), OTM Academy Doctoral Consortium, RDDS (International Workshop on Reliability in Decentralized Distributed Systems), SWWS (IFIP WG 2.12 and WG 12.4 International Workshop on Web Semantics), and SeBGIS (International Workshop on Semantic-based Geographical Information Systems).

OTM 2006 Federated Workshops are proudly supported by CNRS (Centre National de la Recherche Scientifique, France), the City of Montpellier (France), Ecole Polytechnique Universitaire de Montpellier, Université de Montpellier II (UM2), Laboratoire d’Informatique de Robotique et de Microélectronique de Montpellier (LIRMM), RMIT University (School of Computer Science and Information Technology), and Vrije Universiteit Brussel (Department of Computer Science).

Executive Committee

OTM 2006 General Co-chairs:

Robert Meersman (Vrije Universiteit Brussel, Belgium), Zahir Tari (RMIT University, Australia), and Pilar Herrero (Universidad Politécnica de Madrid, Spain).

AWeSOMe 2006 PC Co-chairs:

Daniel Grosu (Wayne State University, USA), Pilar Herrero (Universidad Politécnica de Madrid, Spain), Gonzalo Médez (Universidad Complutense de Madrid, Spain), and Marta Sabou (The Open University, UK).

- CAMS 2006 PC Co-chairs: Annika Hinze (University of Waikato, New Zealand) and George Buchanan (University of Wales Swansea, UK).
- COMINF 2006 PC Co-chairs: Aldo de Moor (Vrije Universiteit Brussel, Belgium) and Michael Gurstein (Community Informatics Research Network, Canada).
- IS 2006 PC Co-chairs: Mário Freire (University of Beira Interior, Portugal), Simão Melo de Sousa (University of Beira Interior, Portugal), and Vitor Santos (Microsoft Lisbon, Portugal).
- KSinBIT 2006 PC Co-chairs: Maja Hadzic (Curtin University of Technology, Australia), Bart De Moor (Katholieke Universiteit Leuven, Belgium), Yves Moreau (Katholieke Universiteit Leuven, Belgium), and Arek Kasprzyk (European Bioinformatics Institute, UK).
- MIOS-CIAO 2006 PC Co-chairs: Antonia Albani (University of Augsburg, Germany) and Jan L.G. Dietz (Delft University of Technology, Netherlands).
- MONET 2006 PC Co-chairs: Fernando Ferri (National Research Council, Italy), Maurizio Rafanelli (National Research Council, Italy) and Arianna D'Ulizia (National Research Council, Italy).
- OnToContent 2006 PC Co-chairs: Mustafa Jarrar (Vrije Universiteit Brussel, Belgium), Claude Ostyn (IEEE-LTSC, USA), Werner Ceusters (University of Buffalo, USA), and Andreas Persidis (Biovista, Greece).
- ORM 2006 PC Co-chairs: Terry Halpin (Neumont University, USA) and Robert Meersman (Vrije Universiteit Brussel, Belgium).
- PerSys 2006 PC Co-chairs: Skevos Evripidou (University of Cyprus, Cyprus) and Roy Campbell (University of Illinois at Urbana-Champaign, USA).
- OTM 2006 Academy Doctoral Consortium PC Co-chairs: Jan Dietz, OTM Academy Dean (Tu Delft, Netherlands), Antonia Albani (University of Augsburg, Germany), Gábor Nagypál (Forschungszentrum Informatik - FZI, Germany) and Johannes Maria Zaha (Queensland University of Technology, Australia).

RDDS 2006 PC Co-chairs:	Eiko Yoneki (University of Cambridge, UK) and Pascal Felber (Université de Neuchâtel, Switzerland).
SWWS 2006 PC Co-chairs:	Katia Sycara (Carnegie Mellon University, USA), Elizabeth Chang (Curtin University of Technology, Australia), Ernesto Damiani (Milan University, Italy), Mustafa Jarrar (Vrije Universiteit Brussel, Belgium), and Tharam Dillon (University of Technology Sydney, Australia) .
SeBGIS 2006 PC Co-chair:	Esteban Zimány (Université Libre de Bruxelles, Belgium).
Publication Co-chairs:	Kwong Yuen Lai (RMIT University) and Peter Dimopoulos (RMIT University).
Local Organizing Chair:	Zohra Bellahsene (University of Montpellier II, France).
Publicity Chair:	Gonzalo Méndez (Universidad Complutense de Madrid, Spain).
Secretariat:	Ana-Cecilia Martinez Barbosa, Jan Demey, and Daniel Meersman.

AWeSOMe (Agents, Web Services and Ontologies Merging) 2006 Program Committee

José Luis Bosque	Marí Pérez
Juan A. Botía Blaya	Ronald Poell
Liliana Cabral	Debbie Richards
Isaac Chao	Víctor Robles
Adam Cheyer	Paul Roe
Ian Dickinson	Manuel Salvadores
John Domingue	Alberto Sánchez
Antonio Garcia Dopico	Weisong Shi
Jorge Gómez	Marius-Calin Silaghi
Dominic Greenwood	Henry Tirri
Jingshan Huang	Santtu Toivonen
Margaret Lyell	Rainer Unland
Dan Marinescu	Chris van Aart
Gregorio Martinez	Sander van Splunter
Michael Maximilien	Julita Vassileva
Mauricio Paletta	Niek Wijngaards
Juan Pavón	Cheng-Zhong Xu
José Mariía Peña	

CAMS (Context-Aware Mobile Systems) 2006 Program Committee

Susanne Boll
Dan Chalmers
Keith Cheverst
Trevor Collins
Gill Dobbie
Tiong Goh
John Grundy
Reto Krummenacher

Diane Lingrand
Dave Nichols
Michel Scholl
Goce Trajcevski
Mark van Setten
Agnes Voisard

COMINF (Community Informatics) 2006 Program Committee

Mark Aakhus
Mark Ackerman
Anjo Anjewierden
Michael Bieber
Andy Bytheway
Stijn Christiaens
Tanguy Coenen
Fiorella De Cindio
Peter Day
Dan Dixon
Lilia Efimova
Hamid Ekbia
Marcus Foth

Mark Gaved
Tom Horan
Driss Ketani
Rolf Kleef
Ulrike Lechner
Peter Mambrey
Dave Newman
Jack Park
Larry Stillman
Beverly Trayner
Bartel Van de Walle
Brian Whitworth

IS (Information Security) 2006 Program Committee

Andre Adelsbach
Manuel Bernardo Barbosa
Carlo Blundo
Fernando Boavida
Thierry Brouard
Ilyoung Chong
Nathan Clarke
Miguel P. Correia
Gwenaël Dërr
Paul Dowland
Mahmoud T. El-Hadidi

Steven Furnell
Michael Gertz
Javier Herranz
Sushil Jajodia
Lech J. Janczewski
Hyun-Kook Kahng
Stamatios Kartalopoulos
Kwok-Yan Lam
Benoit Libert
William List
Henrique S. Mamede

Evangelos Markatos
 Sjouke Mauw
 Natalie Miloslavskaya
 José Pina Miranda
 Edmundo Monteiro
 Yi Mu
 Nuno Ferreira Neves
 Maria Papadaki
 Manuela Pereira
 Hartmut Pohl
 Carlos Ribeiro

Henrique Santos
 Ryoichi Sasaki
 Sung-Won Sohn
 K.P. Subbalakshmi
 Stephanie Teufel
 José Esgalhado Valen
 Serge Vaudenay
 Jozef Vyskoc
 Paulo Veríssimo
 André Zúquete

IS (Information Security) 2006 Additional Referees

Anyi Liu
 Chao Yao
 Filipe Caldeira
 João Abrunhosa
 Jorge Granjal
 José Carlos Bacelar Almeida
 Lingyu Wang
 Nuno Garcia

Pedro Inácio
 Romeu Silva
 Sérgio Nunes
 Shiping Chen
 Sotiris Ioannidis
 Stéahine Cuachie

KSinBIT (Knowledge Systems in Bioinformatics) 2006 Program Committee

Robert Meersman
 Werner Ceusters
 Georges De Moor
 Elizabeth Chang
 Peter Dawyndt
 Jan Van den Bussche
 Antoon Goderis
 Paolo Romano
 Marie-Dominique Devignes
 Bert Coessens
 Mark Wilkinson
 Katy Wolstencroft
 Peter Li
 Robert Stevens

Carole Goble
 Phillip Lord
 Chris Wroe
 Michael Bada
 Ilkay Altintas
 Stephen Potter
 Vasa Curcin
 Armin Haller
 Eyal Oren
 M. Scott Marshall
 Marco Roos
 Iwei Yeh

MIOS+CIAO (Inter-organizational Systems) 2006 Program Committee

Wil van der Aalst
Paer Agerfalk
Antonia Albani
Bernhard Bauer
Emmanuel delaHostria
Jan L.G. Dietz
Johann Eder
Joaquim Filipe
Rony G. Flatscher
Kees van Hee
Jan Hoogervorst
Christian Huemer
Zahir Irani
Peter Loos
Graham McLeod
Arturo Molina

Aldo de Moor
Maira Norrie
Maria Orlowska
Erik Proper
Gil Regev
Dewald Roode
Pnina Soffer
Arne Solvberg
Jose Tribolet
Klaus Turowski
Vijay K. Vaishnavi
Rene Wagenaar
Christian Winnewisser
Robert Winter
Johannes Maria Zaha

MONET (MOBILE and NETWORKING Technologies for social applications) 2006 Program Committee

Russell Beale
Tiziana Catarci
Richard Chbeir
Karin Coninx
Peter Dimopoulos
Juan De Lara
Anna Formica
Patrizia Grifoni
Otthein Herzog
Irina Kondratova
Steve Marsh

Rebecca Montanari
Michele Missikoff
Nuria Oliver
Marco Padula
Andrew Phippen
Tommo Reti
Tim Strayer
Henri Ter Hofte
Riccardo Torlone
Mikael Wiberg

OnToContent (Ontology content and evaluation in Enterprise) 2006 Program Committee

Adil Hameed
Alain Leger
Aldo Gangemi
Andre Valente

Andrew Stranieri
Avigdor Gal
Barry Smith
Bill Andersen

Bob Colomb
 Christiane Fellbaum
 Christopher Brewster
 Ernesto Damiani
 Fausto Giunchiglia
 Francesco Danza
 Francky Trichet
 Giancarlo Guizzardi
 Giorgos Stamou
 Hans Akkermans
 Jeff Pan
 Jens Lemcke
 John Sowa
 Joost Breuker
 Karl Stroetmann
 Kewen Wang
 Luk Vervenne

Miguel-Angel Sicilia
 Mohand-Said Hacid
 Nikolay Mehandjiev
 Paolo Bouquet
 Paul Piwek
 Robert Meersman
 Robert Tolksdorf
 Sergio Tessaris
 Silvie Spreeuwenberg
 Simon White
 Stephen McGibbon
 Theo Mensen
 Yannick Legre
 Yannis Charalabidis
 Yaser Bishr

ORM (Object-Role Modeling) 2006 Program Committee

Guido Bakema
 Herman Balsters
 Linda Bird
 Anthony Bloesch
 Scott Becker
 Peter Bollen
 Andy Carver
 Dave Cuyler
 Necito dela Cruz
 Aldo de Moor
 Olga De Troyer
 Jan Dietz
 David Embley
 Ken Evans
 Gordon Everest
 Henri Habrias
 Pat Hallock

Terry Halpin
 Hank Hermans
 Stijn Hoppenbrouwers
 Mike Jackson
 Mustafa Jarrar
 Alberto Laender
 Inge Lemmens
 Robert Meersman
 Tony Morgan
 Maurice Nijssen
 Sjir Nijssen
 Erik Proper
 Peter Spyns
 Sten Sundblad
 Theo van der Weide
 Gerd Wagner

PerSys (Pervasive Systems) 2006 Program Committee

Jalal Al-Muhtadi
 Jean Bacon

Christain Becker
 Roy Campbell

XVIII Organization

Adrian David Cheok
Hakan Duman
Hesham El-Rewini
Skevos Evripidou
Hans Gellersen
Markus Huebscher
Yunhao Liu
Wallid Najjar

Das Sajal
George Samaras
Anja Schanzenberger
Behrooz Shirazi
Sotirios Terzis
Gregory Yovanof
Arkady Zaslavsky

OTM Academy (International Doctoral Symposium) 2006 Program Committee

Antonia Albani
Domenico Beneventano
Sonia Bergamaschi
Jaime Delgado
Jan Dietz

Johann Eder
Maria Orłowska
Gábor Nagypál
Johannes Maria Zaha

RDDS (Reliability in Decentralized Distributed Systems) 2006 Program Committee

Licia Capra
Mariano Cilia
Vittorio Cortellessa
Simon Courtenage
Patrick Eugster
Ludger Fiege
Maria Gradinariu
Eli Katsiri
Michael Kounavis
Marco Mamei

Jon Munson
Maziar Nekovee
Andrea Passarella
Peter Pietzuch
Matthieu Roy
Francois Taïani
Niki Trigoni
Einar Vollset

SWWS (Web Semantics) 2006 Program Committee

Aldo Gangemi
Amit Sheth
André Valente
Andrew Stranieri
Avigdor Gal
Carole Goble
Carole Hafner

Cecilia Magnusson Sjöberg
Chris Bussler
David Bell
Elisa Bertino
Elizabeth Chang
Enrico Franconi
Ernesto Damiani

Feng Ling
 Frank van Harmelen
 Giancarlo Guizzardi
 Grigoris Antoniou
 Guirau de Lame
 Hai Zhuge
 Jaime Delgado
 Jaiwei Han
 John Debenham
 John Mylopoulos
 Joost Breuker
 Jos Lehmann
 Katia Sycara
 Kokou Yetongnon
 Layman Allen
 Leonardo Lesmo
 Ling Liu
 Lizhu Zhou
 Lotfi Zadeh
 Manfred Hauswirth
 Mariano Lopez
 Masood Nikvesh
 Mihaela Ulieru
 Mohand-Said Hacid

Mukesh Mohania
 Mustafa Jarrar
 Nicola Guarino
 Peter Spyns
 Pièree Yves Schobbens
 Qing Li
 Radboud Winkels
 Ramasamy Uthurusamy
 Richard Benjamins
 Rita Temmerman
 Robert Meersman
 Robert Tolksdorf
 Said Tabet
 Stefan Decker
 Susan Urban
 Tharam Dillon
 Trevor Bench-Capon
 Usuama Fayed
 Valentina Tamma
 Wil van der Aalst
 York Sure
 Zahir Tari

SeBGIS (Semantic-based Geographical Information Systems) 2006 Program Committee

Gennady Adrienko
 Yvan Bédard
 David Bennett
 Michela Bertolotto
 Roland Billen
 Alex Borgida
 Bénédicte Buch
 Christophe Claramunt
 Eliseo Clementini
 Nadine Cullot
 Fernando Ferri
 Anders Friis-Christensen
 Antony Galton

Marinos Kavouras
 Werner Kuhn
 Robert Laurini
 Sergei Levashkin
 Thérèse Libourel
 Peter van Oosterom
 Dimitris Papadias
 Maurizio Rafanelli
 Simonas Saltenis
 Emmanuel Stefanakis
 Nectaria Tryfona
 Stephan Winter

Table of Contents – Part I

Posters of the 2006 CoopIS (Cooperative Information Systems) International Conference

Specifying Instance Correspondence in Collaborative Business Processes	1
<i>Xiaohui Zhao, Chengfei Liu, Yun Yang</i>	
Middleware Support for Monitoring: Privacy and Audit in Healthcare Applications	3
<i>Brian Shand, Jem Rashbass</i>	
A Probabilistic Approach to Reduce the Number of Deadline Violations and the Tardiness of Workflows	5
<i>Johann Eder, Hannes Eichner, Horst Pichler</i>	
A Unified Model for Information Integration	8
<i>Ali Kiani, Nematollaah Shiri</i>	
Service Composition and Deployment for a Smart Items Infrastructure	10
<i>Holger Ziekow, Artin Avanes, Christof Bornhövd</i>	
Synthetizing RDF Query Answers from Multiple Documents	12
<i>Adrian Tanasescu, Mohand-Said Hacid</i>	
Establishing Agile Partnerships in Open Environments: Extended Abstract	15
<i>I.D. Stalker, M. Carpenter, N.D. Mehandjiev</i>	

Posters of the 2006 DOA (Distributed Objects and Applications) International Conference

Distributed Adaptation Reasoning for a Mobility and Adaptation Enabling Middleware	17
<i>Nearchos Paspallis, George A. Papadopoulos</i>	
Scheduling of Composite Web Services	19
<i>Dmytro Dyachuk, Ralph Deters</i>	
Handling and Resolving Conflicts in Real Time Mobile Collaboration	21
<i>Sandy Citro, Jim McGovern, Caspar Ryan</i>	

A Configurable Event Correlation Architecture for Adaptive J2EE Applications	23
<i>Yan Liu, Ian Gorton, Khanh Vinh Le</i>	
Autonomous Deployment and Reconfiguration of Component-Based Applications in Open Distributed Environments	26
<i>Jérémy Dubus, Philippe Merle</i>	
Dynamic Integration of Peer-to-Peer Services into a CORBA-Compliant Middleware	28
<i>Rüdiger Kapitza, Udo Bartlang, Holger Schmidt, Franz J. Hauck</i>	
Automated Deployment of Enterprise Systems in Large-Scale Environments	30
<i>Takoua Abdellatif, Didier Hoareau, Yves Mahéo</i>	
Supporting Reconfigurable Object Distribution for Customizable Web Applications	32
<i>Po-Hao Chang, Gul Agha</i>	
Towards the Definition and Validation of Coupling Metrics for Predicting Maintainability in Service-Oriented Designs	34
<i>Mikhail Pereplechikov, Caspar Ryan, Keith Frampton</i>	
Posters of the 2006 ODBASE (Ontologies, Databases, and Applications of Semantics) International Conference	
R&D Project Management with ODESeW	36
<i>Asunción Gómez-Pérez, Angel López-Cima, María del Carmen Suárez-Figueroa, Oscar Corcho</i>	
Heavyweight Ontology Engineering	38
<i>Frédéric Fürst, Francky Trichet</i>	
Semantic Data Integration in a Newspaper Content Management System	40
<i>A. Abelló, R. García, R. Gil, M. Oliva, F. Perdrix</i>	
Change Detection in Ontologies Using DAG Comparison	42
<i>Johann Eder, Karl Wiggisser</i>	
Filtering Internet News Feeds Using Ad-Hoc Ontologies	44
<i>Lars Bröcker, Stefan Paal</i>	

Mediation as Recommendation: An Approach to Design Mediators for Object Catalogs	46
<i>Daniela F. Brauner, Marco A. Casanova, Ruy L. Milidiú</i>	

Benchmarking Data Schemes of Ontology Based Databases	48
<i>Hondjack Dehainsala, Guy Pierra, Ladjel Bellatreche</i>	

Posters of the 2006 GADA (Grid Computing, High Performance and Distributed Applications) International Conference

Dynamic Load Balancing for Online Games Using Predefined Area Information	50
<i>Beob Kyun Kim, Dong Un An, Seung Jong Chung</i>	

Seamless Integration of Generic Bulk Operations in Grid Applications	52
<i>Stephan Hirmer, Hartmut Kaiser, Andre Merzky, Andrei Hutanu, Gabrielle Allen</i>	

An Efficient Search Scheme Using Self-organizing Hierarchical Ring in Unstructured Peer-to-Peer Systems	55
<i>Saeyoung Han, Jaeeui Sohn, Sungyong Park</i>	

Workshop on Agents, Web Services and Ontologies Merging (AweSOMe)

AWeSOMe 2006 PC Co-chairs' Message	57
--	----

Invited Talk

3-Level Service Composition and Cashew: A Model for Orchestration and Choreography in Semantic Web Services	58
<i>Barry Norton, Carlos Pedrinaci</i>	

Combing Agent and Web Service Technology

Learning from an Active Participation in the Battlefield: A New Web Service Human-Based Approach	68
<i>Mauricio Paletta, Pilar Herrero</i>	

A Collaborative Awareness Specification to Cover Load Balancing Delivery in CSCW Grid Applications	78
<i>Pilar Herrero, José Luis Bosque, Manuel Salvadores, María S. Pérez</i>	

An Agent System for Automated Web Service Composition
and Invocation 90
In-Cheol Kim, Hoon Jin

Web Services in E-Commerce and Business

An Auction-Based Semantic Service Discovery Model for E-Commerce
Applications 97
Vedran Podobnik, Krunoslav Trzec, Gordana Jezic

Implementation of Business Process Requiring User Interaction 107
Guillermo López, Valeria de Castro, Esperanza Marcos

A Petri Net Based Approach for Reliability Prediction
of Web Services 116
Duhang Zhong, Zhichang Qi

Modeling and Re-using Ontologies

Facilitating Ontology (Re)use by Means of a Categorization
Framework 126
Peter De Baer, Koen Kerremans, Rita Temmerman

Agent-Grid Integration Ontology 136
*Frederic Duvert, Clement Jonquet, Pascal Dugenie,
Stefano A. Cerri*

Workshop on Community Informatics (COMINF)

COMINF 2006 PC Co-chairs' Message 147

Community Informatics Foundations

Community Informatics and Human Development 149
William McIver Jr.

Communications Breakdown: Revisiting the Question of Information
and Its Significance for Community Informatics Projects 160
William Tibben

More Than Wires, Pipes and Ducts: Some Lessons from Grassroots
Networked Communities and Master-Planned Neighbourhoods 171
Mark Gaved, Marcus Foth

Capturing Community Meaning

Community-Driven Ontology Evolution Based on Folksonomies	181
<i>Domenico Gendarmi, Filippo Lanubile</i>	
Knowledge Sharing over Social Networking Systems: Architecture, Usage Patterns and Their Application	189
<i>Tanguy Coenen, Dirk Kenis, Céline Van Damme, Eiblin Matthys</i>	
Metadata Mechanisms: From Ontology to Folksonomy ... and Back	199
<i>Stijn Christiaens</i>	

Improving Community Communications

Virtual Individual Networks: A Case Study	208
<i>Licia Calvi</i>	
Agent Community Support for Crisis-Response Organizations	218
<i>Hans Weigand</i>	
Aggregating Information and Enforcing Awareness Across Communities with the Dynamo RSS Feeds Creation Engine: Preliminary Report	227
<i>Fiorella De Cindio, Giacomo Fiumara, Massimo Marchi, Alessandro Provetti, Laura Ripamonti, Leonardo Sonnante</i>	

Analyzing and Designing Community IS

Incorporating Indigenous World Views in Community Informatics	237
<i>Larry Stillman, Barbara Craig</i>	
Towards a Theory of Online Social Rights	247
<i>Brian Whitworth, Aldo de Moor, Tong Liu</i>	
Requirements Determination in a Community Informatics Project: An Activity Theory Approach	257
<i>Melih Kirlidog</i>	
Developing Enterprise Sponsored Virtual Communities: The Case of a SME's Knowledge Community	269
<i>António Lucas Soares, Dora Simões, Manuel Silva, Ricardo Madureira</i>	

Community Evaluation and Assessment Methodologies

Understanding Weblog Communities Through Digital Traces: A Framework, a Tool and an Example	279
<i>Anjo Anjewierden, Lilia Efimova</i>	
Evaluating the Informational and Social Aspects of Participation in Online Communities	290
<i>Emilie Marquois-Ogez, Cécile Bothorel</i>	
An Approach to the Assessment of Applied Information Systems with Particular Application to Community Based Systems	301
<i>Driss Kettani, Michael Gurstein, Bernard Moulin, Asmae El Mahdi</i>	

Plenary Discussion: Towards a Socio-technical Research Agenda for Community Informatics

Workshop on Information Security (IS)

IS 2006 PC Co-chairs' Message	311
---	-----

Multimedia Security

An Implementation of a Trusted and Secure DRM Architecture	312
<i>Víctor Torres, Jaime Delgado, Silvia Llorente</i>	
Using Image Steganography for Decryptor Distribution	322
<i>T. Morkel, J.H.P. Eloff, M.S. Olivier</i>	
An Efficient Dispute Resolving Method for Digital Images	331
<i>Yunho Lee, Heasuk Jo, Seungjoo Kim, Dongho Won</i>	
Use of SAML for Single Sign-On Access to Multimedia Contents in a Peer-to-Peer Network	342
<i>Rubén Barrio, Xavier Perramon, Jaime Delgado</i>	

RFID Security

EMAP: An Efficient Mutual-Authentication Protocol for Low-Cost RFID Tags	352
<i>Pedro Peris-Lopez, Julio Cesar Hernandez-Castro, Juan M. Estevez-Tapiador, Arturo Ribagorda</i>	

Secure EPCglobal Class-1 Gen-2 RFID System Against Security and Privacy Problems	362
<i>Kyoungh Hyun Kim, Eun Young Choi, Su Mi Lee, Dong Hoon Lee</i>	

A Case Against Currently Used Hash Functions in RFID Protocols	372
<i>Martin Feldhofer, Christian Rechberger</i>	

Network Security

An Efficient ID-Based Delegation Network	382
<i>Taek-Young Youn, Young-Ho Park, Chang Han Kim, Jongin Lim</i>	

Enabling Practical IPsec Authentication for the Internet	392
<i>Pedro J. Muñoz Merino, Alberto García Martínez, Mario Muñoz Organero, Carlos Delgado Kloos</i>	

Preamble Encryption Mechanism for Enhanced Privacy in Ethernet Passive Optical Networks	404
<i>Pedro R.M. Inácio, Marek Hajduczenia, Mário M. Freire, Henrique J.A. da Silva, Paulo P. Monteiro</i>	

SMARTCOP – A Smart Card Based Access Control for the Protection of Network Security Components	415
<i>Joaquín García-Alfaro, Sergio Castillo, Jordi Castellà-Roca, Guillermo Navarro, Joan Borrell</i>	

Cryptographic Algorithms and Protocols

On the Existence of Related-Key Oracles in Cryptosystems Based on Block Ciphers	425
<i>Ermaliza Razali, Raphael Chung Wei Phan</i>	

New Key Generation Algorithms for the XTR Cryptosystem	439
<i>Maciej Grześkowiak</i>	

Public-Key Encryption from ID-Based Encryption Without One-Time Signature	450
<i>Chik How Tan</i>	

Solving Bao's Colluding Attack in Wang's Fair Payment Protocol	460
<i>M. Magdalena Payeras-Capellà, Josep L. Ferrer Gomila, Llorenç Huguet Rotger</i>	

Biometrics for Security

An Efficient Algorithm for Fingercod-Based Biometric Identification	469
<i>Hong-Wei Sun, Kwok-Yan Lam, Ming Gu, Jia-Guang Sun</i>	
Robustness of Biometric Gait Authentication Against Impersonation Attack	479
<i>Davrondzhon Gafurov, Einar Snekkenes, Tor Erik Buvarp</i>	
From Features Extraction to Strong Security in Mobile Environment: A New Hybrid System	489
<i>Stéphane Cauchie, Thierry Brouard, Hubert Cardot</i>	

Access Control and Smart Card Technology

Improving the Dynamic ID-Based Remote Mutual Authentication Scheme	499
<i>Eun-Jun Yoon, Kee-Young Yoo</i>	
Security Enhancement of a Remote User Authentication Scheme Using Smart Cards	508
<i>Youngsook Lee, Junghyun Nam, Dongho Won</i>	
Information Leakage and Capability Forgery in a Capability-Based Operating System Kernel	517
<i>Dan Mossop, Ronald Pose</i>	
Reverse Engineering of Embedded Software Using Syntactic Pattern Recognition	527
<i>Mike Fournigault, Pierre-Yvan Liardet, Yannick Teglia, Alain Trémeau, Frédérique Robert-Inacio</i>	

Risk Analysis and Business Continuity

Disaster Coverable PKI Model Utilizing the Existing PKI Structure	537
<i>Bo Man Kim, Kyu Young Choi, Dong Hoon Lee</i>	
A Corporate Capital Protection and Assurance Model	546
<i>Colette Reekie, Basie von Solms</i>	
Quantitative Evaluation of Systems with Security Patterns Using a Fuzzy Approach	554
<i>Spyros T. Halkidis, Alexander Chatzigeorgiou, George Stephanides</i>	

Managing Critical Information Infrastructure Security Compliance: A Standard Based Approach Using ISO/IEC 17799 and 27001	565
<i>Wipul Jayawickrama</i>	

Mobile and Wireless Network Security

A Composite Key Management Scheme for Mobile Ad Hoc Networks	575
<i>Yingfang Fu, Jingsha He, Guorui Li</i>	
Adaptive Algorithms to Enhance Routing and Security for Wireless PAN Mesh Networks	585
<i>Cao Trong Hieu, Tran Thanh Dai, Choong Seon Hong, Jae-Jo Lee</i>	
Secure and Seamless Handoff Scheme for a Wireless LAN System	595
<i>Jaesung Park, Beomjoon Kim, Iksoon Hwang</i>	

Information Security and Service Resilience

A CAPTCHA in the Text Domain	605
<i>Pablo Ximenes, André dos Santos, Marcial Fernandez, Joaquim Celestino Jr.</i>	
Examining the DoS Resistance of HIP	616
<i>Suratose Tritilanunt, Colin Boyd, Ernest Foo, Juan Manuel González Nieto</i>	
Securing Data Accountability in Decentralized Systems	626
<i>Ricardo Corin, David Galindo, Jaap-Henk Hoepman</i>	
Privacy Friendly Information Disclosure	636
<i>Steven Gevers, Bart De Decker</i>	

Workshop on Knowledge Systems in Bioinformatics (KsinBIT)

KSinBIT 2006 PC Co-chairs' Message	647
--	-----

Bioinformatics Ontologies, Terminologies and Knowledge Bases

Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing	648
<i>Adrien Coulet, Malika Smail-Tabbone, Amedeo Napoli, Marie-Dominique Devignes</i>	

Methodologically Designing a Hierarchically Organized
 Concept-Based Terminology Database to Improve Access to Biomedical
 Documentation 658
*Antonio Vaquero, Fernando Sáenz, Francisco Alvarez,
 Manuel de Buenaga*

A Proposal for a Gene Functions Wiki 669
*Robert Hoehndorf, Kay Prüfer, Michael Backhaus, Heinrich Herre,
 Janet Kelso, Frank Loebe, Johann Visagie*

Use of Ontologies in Bioinformatics

Using Semantic Web Tools to Integrate Experimental Measurement
 Data on Our Own Terms 679
M. Scott Marshall, Lennart Post, Marco Roos, Timo M. Breit

Ontology Guided Data Integration for Computational Prioritization
 of Disease Genes 689
*Bert Coessens, Stijn Christiaens, Ruben Verlinden, Yves Moreau,
 Robert Meersman, Bart De Moor*

Bringing Together Structured and Unstructured Sources:
 The OUMSUIS Approach 699
Gayo Diallo, Michel Simonet, Ana Simonet

**Bioinformatics Data Manipulation, Integration
 and Associated Tools**

Reactome – A Knowledgebase of Biological Pathways 710
*Esther Schmidt, Ewan Birney, David Croft, Bernard de Bono,
 Peter D’Eustachio, Marc Gillespie, Gopal Gopinath, Bijay Jassal,
 Suzanna Lewis, Lisa Matthews, Lincoln Stein, Imre Vastrik,
 Guanming Wu*

Structural Similarity Mining in Semi-structured Microarray Data
 for Efficient Storage Construction 720
Jongil Jeong, Dongil Shin, Chulho Cho, Dongkyoo Shin

Modeling and Storing Scientific Protocols 730
Natalia Kwasnikowska, Yi Chen, Zoé Lacroix

A Knuckles-and-Nodes Approach to the Integration
 of Microbiological Resource Data 740
*Bart Van Brabant, Peter Dawyndt, Bernard De Baets,
 Paul De Vos*

Workshop on Modeling Inter-Organizational Systems (MIOS-CIAO)

MIOS-CIAO 2006 PC Co-chairs' Message	751
--	-----

Architecture in Inter-organizational Cooperation

The CrocodileAgent: Research for Efficient Agent-Based Cross-Enterprise Processes	752
<i>Vedran Podobnik, Ana Petric, Gordan Jezic</i>	

Case Study – Automating Direct Banking Customer Service Processes with Service Oriented Architecture	763
<i>Andreas Eberhardt, Oliver Gausmann, Antonia Albani</i>	

An Event-Trigger-Rule Model for Supporting Collaborative Knowledge Sharing Among Distributed Organizations	780
<i>Minsoo Lee, Stanley Y.W. Su, Herman Lam</i>	

Ontology and Project Management

Semantic LBS: Ontological Approach for Enhancing Interoperability in Location Based Services	792
<i>Jong-Woo Kim, Ju-Yeon Kim, Chang-Soo Kim</i>	

Dynamic Consistency Between Value and Coordination Models – Research Issues	802
<i>Lianne Bodenstaff, Andreas Wombacher, Manfred Reichert</i>	

PROMONT – A Project Management Ontology as a Reference for Virtual Project Organizations	813
<i>Sven Abels, Frederik Ahlemann, Axel Hahn, Kevin Hausmann, Jan Strickmann</i>	

SPI Methodology for Virtual Organizations	824
<i>Paula Ventura Martins, Alberto Rodrigues da Silva</i>	

Inter-organizational Business Processes

A Pattern-Knowledge Base Supported Establishment of Inter-organizational Business Processes	834
<i>Alex Norta, Marcel Hendrix, Paul Grefen</i>	

On the Concurrency of Inter-organizational Business Processes	844
<i>Diogo R. Ferreira</i>	

From Inter-organizational to Inter-departmental Collaboration – Using
Multiple Process Levels 854
Daniel Simonovich

An Ontology-Based Scheme Enabling the Modeling of Cooperation
in Business Processes 863
Manuel Noguera, M. Visitación Hurtado, José L. Garrido

**Workshop on Mobile and Networking Technologies
for Social Applications (MONET)**

MONET 2006 PC Co-chairs' Message 873

Social Networks

Exploring Social Context with the Wireless Rope 874
Tom Nicolai, Eiko Yoneki, Nils Behrens, Holger Kenn

Extending Social Networks with Implicit Human-Human Interaction 884
Tim Clerckx, Geert Houben, Kris Luyten, Karin Coninx

A Classification of Trust Systems 894
Sebastian Ries, Jussi Kangasharju, Max Mühlhäuser

**Studies and Methodologies for Mobile
and Networking Technologies**

Solving Ambiguities for Sketch-Based Interaction in Mobile
Environments 904
Danilo Avola, Maria Chiara Caschera, Patrizia Grifoni

Supporting Mobile Activities in a Shared Semantic Ambient 916
Fabio Pittarello, Augusto Celentano

Cultural Interface Design: Global Colors Study 926
Irina Kondratova, Ilia Goldfarb

**Mobile and Networking Technologies in Different
Context**

Multimodal Interactive Systems to Manage Networked Human Work 935
*Giuseppe Fresta, Andrea Marcante, Piero Mussio,
Elisabetta Oliveri, Marco Padula*

SIAPAS: A Case Study on the Use of a GPS-Based Parking System	945
<i>Gonzalo Mendez, Pilar Herrero, Ramon Valladares</i>	

Mobile Social Software for Cultural Heritage Management	955
<i>Yiwei Cao, Satish Narayana Srirama, Mohamed Amine Chatti, Ralf Klamma</i>	

Architecture and Middleware

Middleware Platform for Ubiquitous Location Based Service	965
<i>Jae-Chul Kim, Jai-Ho Lee, Ju-Wan Kim, Jong-Hyun Park</i>	

Multimodal Architectures: Issues and Experiences	974
<i>Giovanni Frattini, Luigi Romano, Vladimiro Scotto di Carlo, Pierpaolo Petriccione, Gianluca Supino, Giuseppe Leone, Ciro Autiero</i>	

MobiSoft: An Agent-Based Middleware for Social-Mobile Applications . . .	984
<i>Steffen Kern, Peter Braun, Wilhelm Rossak</i>	

Innovative Healthcare Services for Nomadic Users	994
<i>Marcello Melgara, Luigi Romano, Fabio Rocca, Alberto Sanna, Daniela Marino, Riccardo Serafin</i>	

Author Index	1003
-------------------------------	-------------

Table of Contents – Part II

Workshop on Ontology Content and Evaluation in Enterprise (OnToContent)

OnToContent 2006 PC Co-chairs' Message	1011
--	------

Ontology Quality and Consensus

Unit Tests for Ontologies	1012
<i>Denny Vrandečić, Aldo Gangemi</i>	

Issues for Robust Consensus Building in P2P Networks (Position Paper)	1021
<i>Abdul-Rahman Mawlood-Yunis, Michael Weiss, Nicola Santoro</i>	

Ontology and Agent Based Model for Software Development Best Practices' Integration in a Knowledge Management System	1028
<i>Nahla Jlaiel, Mohamed Ben Ahmed</i>	

Extracting Ontological Relations of Korean Numeral Classifiers from Semi-structured Resources Using NLP Techniques (Position Paper)	1038
<i>Youngim Jung, Soonhee Hwang, Aesun Yoon, Hyuk-Chul Kwon</i>	

Ontology Construction and eHealth Ontologies

A Transfusion Ontology for Remote Assistance in Emergency Health Care (Position Paper)	1044
<i>Paolo Ceravolo, Ernesto Damiani, Cristiano Fugazza</i>	

Ontological Distance Measures for Information Visualisation on Conceptual Maps	1050
<i>Sylvie Ranwez, Vincent Ranwez, Jean Villerd, Michel Crampes</i>	

The Management and Integration of Biomedical Knowledge: Application in the Health-e-Child Project (Position Paper)	1062
<i>Ernesto Jimenez-Ruiz, Rafael Berlanga, Ismael Sanz, Richard McClatchey, R. Danger, David Manset, Jordi Paraire, Alfonso Rios</i>	

Competence Ontologies

Ontology-Based Systems Dedicated to Human Resources Management: An Application in e-Recruitment	1068
<i>Vladimir Radevski, Francky Trichet</i>	
Towards a Human Resource Development Ontology for Combining Competence Management and Technology-Enhanced Workplace Learning	1078
<i>Andreas Schmidt, Christine Kunzmann</i>	
The eCCO System: An eCompetence Management Tool Based on Semantic Networks	1088
<i>Barbara Pernici, Paolo Locatelli, Clementina Marinoni</i>	
Competency Model in a Semantic Context: Meaningful Competencies (Position Paper)	1100
<i>Stijn Christiaens, Jan De Bo, Ruben Verlinden</i>	
Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition	1107
<i>Mathieu Roche, Yves Kodratoff</i>	

Workshop on Object-Role Modeling(ORM)

ORM 2006 PC Co-chairs' Message	1117
--	------

Modeling Extensions

Part-Whole Relations in Object-Role Models	1118
<i>C. Maria Keet</i>	
Exploring Modelling Strategies in a Meta-modelling Context	1128
<i>P. van Bommel, S.J.B.A. Hoppenbrouwers, H.A. (Erik) Proper, Th.P. van der Weide</i>	
Giving Meaning to Enterprise Architectures: Architecture Principles with ORM and ORC	1138
<i>P. van Bommel, S.J.B.A. Hoppenbrouwers, H.A. (Erik) Proper, Th.P. van der Weide</i>	

Data Warehousing

- Using ORM-Based Models as a Foundation for a Data Quality Firewall
in an Advanced Generation Data Warehouse 1148
Baba Piprani
- Evolution of a Dynamic Multidimensional Denormalization Meta
Model Using Object Role Modeling 1160
Joe Hansen, Necito dela Cruz

Language and Tool Issues

- Fact-Oriented Modeling from a Programming Language Designer's
Perspective 1170
Betsy Pepels, Rinus Plasmeijer, H.A. (Erik) Proper
- Automated Verbalization for ORM 2 1181
Terry Halpin, Matthew Curland
- T-Lex: A Role-Based Ontology Engineering Tool 1191
*Damien Trog, Jan Vereecken, Stijn Christiaens, Pieter De Leenheer,
Robert Meersman*

Dynamic Rules and Processes

- Modeling Dynamic Rules in ORM 1201
Herman Balsters, Andy Carver, Terry Halpin, Tony Morgan
- Some Features of State Machines in ORM 1211
Tony Morgan

The Modeling Process and Instructional Design

- Data Modeling Patterns Using Fully Communication Oriented
Information Modeling (FCO-IM) 1221
Fazat Nur Azizah, Guido Bakema
- Using Fact-Oriented for Instructional Design 1231
Peter Bollen
- Capturing Modeling Processes – Towards the MoDial Modeling
Laboratory 1242
*S.J.B.A. (Stijn) Hoppenbrouwers, L. (Leonie) Lindeman,
H.A. (Erik) Proper*

Workshop on Pervasive Systems (PerSys)

PerSys 2006 PC Co-chairs' Message 1253

Infrastructure for Pervasive Systems

Live and Heterogeneous Migration of Execution Environments
(Position Paper) 1254
Nicolas Geoffray, Gaël Thomas, Bertil Folliot

A Comparative Study Between Soft System Bus and Traditional
Middlewares 1264
Mohammad Reza Selim, Takumi Endo, Yuichi Goto, Jingde Cheng

Service Discovery and Personalization

A Semantic Location Service for Pervasive Grids 1274
Antonio Coronato, Giuseppe De Pietro, Massimo Esposito

A Pragmatic and Pervasive Methodology to Web Service Discovery 1285
Electra Tamani, Paraskevas Evaripidou

Mobile User Personalization with Dynamic Profiles: Time
and Activity 1295
Christoforos Panayiotou, George Samaras

Pervasive Environments

Rethinking the Design of Enriched Environments 1305
*Felipe Aguilera, Andres Neyem, Rosa A. Alarcón, Luis A. Guerrero,
Cesar A. Collazos*

The eHomeConfigurator Tool Suite 1315
Ulrich Norbistrath, Christof Mosler, Ibrahim Armac

Mobile Phone + RFID Technology = New Mobile Convergence Toward
Ubiquitous Computing Environment 1325
Taesu Cheong, Marie Kim

Clicky: Input in Pervasive Systems 1337
Andrew Weiler, Jeffrey Naisbitt, Roy Campbell

Security and Privacy

- A User-Centric Privacy Framework for Pervasive Environments 1347
*Susana Alcalde Bagüés, Andreas Zeidler,
 Carlos Fernandez Valdivielso, Ignacio R. Matias*
- An XML-Based Security Architecture for Integrating Single Sign-On
 and Rule-Based Access Control in Mobile and Ubiquitous Web
 Environments 1357
Jongil Jeong, Dongil Shin, Dongkyoo Shin
- Architecture Framework for Device Single Sign On in Personal Area
 Networks 1367
*Appadodharana Chandershekarapuram, Dimitrios Vogiatzis,
 Spyridon Vassilaras, Gregory S. Yovanof*

Wireless and Sensor Networks

- Applying the Effective Distance to Location-Dependent Data 1380
JeHyok Ryu, Uğur Çetintemel
- Scalable Context Simulation for Mobile Applications 1391
Y. Liu, M.J. O’Grady, G.M.P. O’Hare

Doctoral Consortium

- OTM 2006 Academy Doctoral Consortium PC Co-chairs’ Message 1401

Enterprise Systems

- Multidimensional Effort Prediction for ERP System Implementation 1402
Torben Hansen
- A Decision-Making Oriented Model for Corporate Mobile Business
 Applications 1409
Daniel Simonovich
- Applying Architecture and Ontology to the Splitting and Allying
 of Enterprises: Problem Definition and Research Approach 1419
Martin Op ’t Land

System Development

Profile-Based Online Data Delivery	1429
<i>Haggai Roitman</i>	
An Incremental Approach to Enhance the Accuracy of Internet Routing	1439
<i>Ines Feki</i>	
Ontology-Based Change Management of Composite Services	1450
<i>Linda Terlouw</i>	
Research on Ontology-Driven Information Retrieval	1460
<i>Stein L. Tomassen</i>	

Workshop on Reliability in Decentralized Distributed Systems (RDDS)

RDDS 2006 PC Co-chairs' Message	1469
---	------

P2P

Core Persistence in Peer-to-Peer Systems: Relating Size to Lifetime	1470
<i>Vincent Gramoli, Anne-Marie Kermarrec, Achour Mostefaoui, Michel Raynal, Bruno Sericola</i>	
Implementing Chord for HP2P Network	1480
<i>Yang Cao, Zhenhua Duan, Jian-Jun Qi, Zhuo Peng, Ertao Lv</i>	
A Generic Self-repair Approach for Overlays	1490
<i>Barry Porter, Geoff Coulson, François Taïani</i>	
Providing Reliable IPv6 Access Service Based on Overlay Network	1500
<i>Xiaoxiang Leng, Jun Bi, Miao Zhang</i>	

Distributed Algorithms

Adaptive Voting for Balancing Data Integrity with Availability	1510
<i>Johannes Osrail, Lorenz Frohofer, Matthias Gladt, Karl M. Goeschka</i>	
Efficient Epidemic Multicast in Heterogeneous Networks	1520
<i>José Pereira, Rui Oliveira, Luís Rodrigues</i>	

Decentralized Resources Management for Grid	1530
<i>Thibault Bernard, Alain Bui, Olivier Flauzac, Cyril Rabat</i>	

Reliability Evaluation

Reliability Evaluation of Group Service Providers in Mobile Ad-Hoc Networks	1540
<i>Joos-Hendrik Böse, Andreas Thaler</i>	

Traceability and Timeliness in Messaging Middleware.....	1551
<i>Brian Shand, Jem Rashbass</i>	

Transaction Manager Failover: A Case Study Using JBOSS Application Server	1555
<i>A.I. Kistijantoro, G. Morgan, S.K. Shrivastava</i>	

Workshop on Semantic-Based Geographical Information Systems (SeBGIS)

SeBGIS 2006 PC Co-chairs' Message	1565
---	------

GIS Integration and GRID

Matching Schemas for Geographical Information Systems Using Semantic Information.....	1566
<i>Christoph Quix, Lemonia Ragia, Linlin Cai, Tian Gan</i>	

Towards a Context Ontology for Geospatial Data Integration	1576
<i>Damires Souza, Ana Carolina Salgado, Patricia Tedesco</i>	

Spatial Data Access Patterns in Semantic Grid Environment	1586
<i>Vikram Sorathia, Anutosh Maitra</i>	

Spatial Data Warehouses

GeWolap: A Web Based Spatial OLAP Proposal	1596
<i>Sandro Bimonte, Pascal Wehrle, Anne Tchounikine, Maryvonne Miquel</i>	

Affordable Web-Based Spatio-temporal Applications for Ad-Hoc Decisions	1606
<i>Vera Hernández Ernst</i>	

Requirements Specification and Conceptual Modeling for Spatial Data
Warehouses 1616
Elzbieta Malinowski, Esteban Zimányi

Spatio-temporal GIS

A Classification of Spatio-temporal Entities Based on Their Location
in Space-Time 1626
Thomas Bittner, Maureen Donnelly

Efficient Storage of Interactions Between Multiple Moving Point
Objects 1636
*Nico Van de Weghe, Frank Witlox, Anthony G. Cohn, Tijs Neutens,
Philippe De Maeyer*

Implementing Conceptual Spatio-temporal Schemas
in Object-Relational DBMSs 1648
Esteban Zimányi, Mohammed Minout

Semantic Similarity

A Semantic Similarity Model for Mapping Between Evolving Geospatial
Data Cubes 1658
Mohamed Bakillah, Mir Abolfazl Mostafavi, Yvan Bédard

Query Approximation by Semantic Similarity in GeoPQL 1670
Fernando Ferri, Anna Formica, Patrizia Grifoni, Maurizio Rafanelli

Sim-DL: Towards a Semantic Similarity Measurement Theory
for the Description Logic $\mathcal{ALCN}\mathcal{R}$ in Geographic Information
Retrieval 1681
Krzysztof Janowicz

Enhancing Usability

A Primer of Geographic Databases Based on Chorems 1693
Robert Laurini, Françoise Milleret-Raffort, Karla Lopez

A Service to Customize the Structure of a Geographic Dataset 1703
Sandrine Balley, Bénédicte Bucher, Thérèse Libourel

Using a Semantic Approach for a Cataloguing Service 1712
*Paul Boisson, Stéphane Clerc, Jean-Christophe Desconnets,
Thérèse Libourel*

IFIP WG 2.12 and WG 12.4 International Workshop on Web Semantics (SWWS)

SWWS 2006 PC Co-chairs' Message	1723
---------------------------------------	------

Security, Risk and Privacy for the Semantic Web

Reputation Ontology for Reputation Systems	1724
<i>Elizabeth Chang, Farookh Khadeer Hussain, Tharam Dillon</i>	

Rule-Based Access Control for Social Networks	1734
<i>Barbara Carminati, Elena Ferrari, Andrea Perego</i>	

An OWL Copyright Ontology for Semantic Digital Rights Management	1745
<i>Roberto García, Rosa Gil</i>	

A Legal Ontology to Support Privacy Preservation in Location-Based Services	1755
<i>Hugo A. Mitre, Ana Isabel González-Tablas, Benjamín Ramos, Arturo Ribagorda</i>	

A Fuzzy Approach to Risk Based Decision Making	1765
<i>Omar Khadeer Hussain, Elizabeth Chang, Farookh Khadeer Hussain, Tharam S. Dillon</i>	

Semantic Web and Querying

Header Metadata Extraction from Semi-structured Documents Using Template Matching	1776
<i>Zewu Huang, Hai Jin, Pingpeng Yuan, Zongfen Han</i>	

Query Terms Abstraction Layers	1786
<i>Stein L. Tomassen, Darijus Strasunskas</i>	

VQS - An Ontology-Based Query System for the SemanticLIFE Digital Memory Project	1796
<i>Hanh Huu Hoang, Amin Andjomshoaa, A Min Tjoa</i>	

Ontologies

Software Design Process Ontology Development	1806
<i>P. Wongthongtham, E. Chang, T. Dillon</i>	

Ontology Views: A Theoretical Perspective 1814
R. Rajugan, Elizabeth Chang, Tharam S. Dillon

OntoExtractor: A Fuzzy-Based Approach to Content
 and Structure-Based Metadata Extraction 1825
Paolo Ceravolo, Ernesto Damiani, Marcello Leida, Marco Viviani

Applications of Semantic Web

Towards Semantic Interoperability of Protein Data Sources 1835
Amandeep S. Sidhu, Tharam S. Dillon, Elizabeth Chang

QP-T: Query Pattern-Based RDB-to-XML Translation 1844
Jinhyung Kim, Dongwon Jeong, Yixin Jing, Doo-Kwon Baik

Concepts for the Semantic Web

A Study on the Use of Multicast Protocol Traffic Overload
 for QCBT 1854
Won-Hyuck Choi, Young-Ho Song

Semantic Granularity for the Semantic Web 1863
*Riccardo Albertoni, Elena Camossi, Monica De Martino,
 Franca Giannini, Marina Monti*

Maximum Rooted Spanning Trees for the Web 1873
Wookey Lee, Seungkil Lim

**Workshop on Context Aware Mobile Systems
 (CAMS)**

CAMS 2006 PC Co-chairs’ Message 1883

Models of Context

An Investigation into a Universal Context Model to Support
 Context-Aware Applications 1884
Jason Pascoe, Helena Rodrigues, César Ariza

A System for Context-Dependent User Modeling 1894
*Petteri Nurmi, Alfons Salden, Sian Lun Lau, Jukka Suomela,
 Michael Sutterer, Jean Millerat, Miquel Martin, Eemil Lagerspetz,
 Remco Poortinga*

Granular Context in Collaborative Mobile Environments 1904
Christoph Dorn, Daniel Schall, Shahram Dustdar

Service Models

Context-Aware Services for Physical Hypermedia Applications	1914
<i>Gustavo Rossi, Silvia Gordillo, Cecilia Challiol, Andrés Fortier</i>	
QoS-Predictions Service: Infrastructural Support for Proactive QoS- and Context-Aware Mobile Services (Position Paper)	1924
<i>Katarzyna Wac, Aart van Halteren, Dimitri Konstantas</i>	

Data Handling

LiteMap: An Ontology Mapping Approach for Mobile Agents’ Context-Awareness	1934
<i>Haïfa Zargayouna, Nejla Amara</i>	
Compressing GPS Data on Mobile Devices	1944
<i>Ryan Lever, Annika Hinze, George Buchanan</i>	
Seamless Service Adaptation in Context-Aware Middleware for U-HealthCare	1948
<i>Eun Jung Ko, Hyung Jik Lee, Jeun Woo Lee</i>	

Users and Uses

Using Laddering and Association Techniques to Develop a User-Friendly Mobile (City) Application	1956
<i>Greet Jans, Licia Calvi</i>	
A Situation-Aware Mobile System to Support Fire Brigades in Emergency Situations	1966
<i>Kris Luyten, Frederik Winters, Karin Coninx, Dries Naudts, Ingrid Moerman</i>	
Context-Based e-Learning Composition and Adaptation	1976
<i>Maria G. Abarca, Rosa A. Alarcon, Rodrigo Barria, David Fuller</i>	

Filtering and Control

Context-Driven Data Filtering: A Methodology	1986
<i>Cristiana Bolchini, Elisa Quintarelli</i>	
A Contextual Attribute-Based Access Control Model	1996
<i>Michael J. Covington, Manoj R. Sastry</i>	

Author Index	2007
-------------------------------	------

Specifying Instance Correspondence in Collaborative Business Processes*

Xiaohui Zhao, Chengfei Liu, and Yun Yang

Centre for Information Technology Research
Faculty of Information and Communication Technologies
Swinburne University of Technology
Melbourne, Victoria, Australia 3122
{xzhaoh, cliu, yyang}@ict.swin.edu.au

In recent years, organisations have been undergoing a thorough transformation towards highly flexible and agile collaborations. Organisations are required to dynamically create and manage collaborative business processes to grasp market opportunities. Different from conventional business processes, a collaborative business process involves multiple parties and their business processes [1]. Thus, complex instance correspondences may exist at both build time and run time. Here, we choose to characterise instance correspondences in terms of cardinality and correlations. Thereby, we can define and represent static and dynamic correspondence when modelling and executing a collaborative business process.

Multiple workflow instantiation was discussed by Dumas and ter Hofstede [2], using UML activity diagrams. Later they extended their work to service interactions [3]. van der Aalst et al. [4] deployed coloured Petri nets to represent multiple workflow cases in workflow patterns, and implemented it in the YAWL system [5]. Zhou, Shi and Ye [6] also studied pattern based modelling for multiple instances of workflow activities. Guabtni and Charoy [7] extended the multiple instantiation patterns and classified multiple workflow instantiation into parallel and iterative instances. Yet, most of these research focus on interaction patterns, and sidestep the instance correspondence issue in collaborative business processes. WS-BPEL (previously BPEL4WS) [8] uses its own correlation set to combine workflow instances, which have same values on specified message fields. However, WS-BPEL defines a business process in terms of a pivot organisation, which results in the lack of interactions beyond neighbouring organisations.

Aiming to address the instance correspondence issue, we propose a method to support instance correspondences from an organisation-oriented view. In our method, cardinality parameters are developed to characterise cardinality relationships between collaborating workflow processes at build time. The four bilateral cardinality relationships, viz., single-to-single, single-to-many, many-to-single and many-to-many, are represented by a pair of unidirectional cardinality relationships, which can be either to-single or to-many. Thereby, the cardinality relationship between workflow processes participating in the same collaboration can be defined from the perspective of a given organisation. Workflow correlation denotes the semantic relation between workflow instances in the same business collaboration. Two or more workflow instances are directly correlated,

* This work is partly supported by the Australian Research Council discovery project DP0557572.

when they “shake hands” during run time interactions. In addition, some participating instances may inherit pre-existing workflow correlations from their correlated instances during run time interactions. This correlation inheritance allows instance correspondence propagate as the collaboration proceeds.

For precise representation, we formalise this method with extended coloured Petri nets, since a coloured Petri net can represent multiple process executions within one net. By doing so, the traditional Petri nets are extended with the proposed cardinality parameters and correlation structures. Besides, auxiliary places are imported to link several separate Petri nets together, and token-generatable transitions are used to indicate the tasks that can create new instances of a business process. Thereby, such an extended Petri net can simulate the behaviours of a self-driven and self-evolving business collaboration scenario. In this Petri net based approach, particular algorithms are given to describe how the individual workflow processes can be assembled into a collaborative business process. Furthermore, algorithms also illustrate how we specify workflow correlations and trace workflow correlations on the fly.

Our research on instance correspondence in collaborative business processes establishes a foundation for advanced instance level applications, such as inter-organisational workflow tracking [9] etc. Our future work is to combine this method with our existing relative workflow framework. This future work is expected to provide a comprehensive solution for collaborative business process applications.

References

1. Zhao, X., Liu, C., and Yang, Y.: An Organisational Perspective on Collaborative Business Processes. In Proceedings of the 3rd International Conference on Business Process Management. Nancy, France. (2005) 17-31.
2. Dumas, M. and ter Hofstede, A.H.M.: UML Activity Diagrams as a Workflow Specification Language. In Proceedings of 4th International Conference on the Unified Modeling Language, Modeling Languages, Concepts, and Tools. Toronto, Canada (2001) 76-90.
3. Barros, A.P., Dumas, M., and ter Hofstede, A.H.M.: Service Interaction Patterns. In Proceedings of Proceedings of the 3rd International Conference on Business Process Management (BPM 2005). Nancy, France (2005) 302-318.
4. van der Aalst, W.M.P., ter Hofstede, A.H.M., Kiepuszewski, B., and Barros, A.P.: Workflow Patterns. *Distributed and Parallel Databases*, 14(1) (2003) 5-51.
5. van der Aalst, W.M.P. and ter Hofstede, A.H.M.: YAWL: Yet Another Workflow Language. *Information Systems*, 30(4) (2005) 245-275.
6. Zhou, J., Shi, M., and Ye, X.: On Pattern-based Modeling for Multiple Instances of Activities in Workflows. In Proceedings of International Workshop on Grid and Cooperative Computing. Hainan, China (2002) 723-736.
7. Guabtni, A. and Charoy, F.: Multiple Instantiation in a Dynamic Workflow Environment. In Proceedings of 16th International Conference on Advanced Information Systems Engineering (CAiSE 2004). Riga, Latvia (2004) 175-188.
8. Andrews, T., Curbera, F., Dholakia, H., Golland, Y., Klein, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., Weerawarana, S.: *Business Process Execution Language for Web Services* (2003)
9. Zhao, X. and Liu, C.: Tracking over Collaborative Business Processes. In Proceedings of the 4th International Conference on Business Process Management. Vienna, Austria (2006) 33-48.

Middleware Support for Monitoring: Privacy and Audit in Healthcare Applications

Brian Shand and Jem Rashbass

Clinical and Biomedical Computing Unit, University of Cambridge
16 Mill Lane, Cambridge CB2 1SB, United Kingdom
{Brian.Shand, jem}@cbcu.cam.ac.uk

Abstract. Healthcare applications have special needs that are not met by current messaging and workflow systems. This paper presents a framework which allows reliable audit of access to medical records, and also provides privacy protection and logistical monitoring information for assessing the effectiveness of medical automation. This is achieved by extending the messaging middleware semantics to support message monitoring tokens and privacy meta-data, together with distributed logging. The same mechanisms provide support for Workflow Management System (WfMS) integration and workflow monitoring; the rich log structure allows the actual workflow usage to be reconstructed as a Petri Net, with greater accuracy than existing techniques. This enables workflow mining and evolution without compromising the privacy of clinical data.

Introduction. Medical applications have complex, sometimes conflicting requirements which need support from their messaging middleware: (1) Audit of access to confidential medical data, (2) Monitoring of component usage, (3) Access control and protection of data privacy and (4) Protection against silent data loss. They also need to support higher-level requirements: quality control of automatic processes (especially for clinical decisions), a high proportion of manual decisions, and monitoring and accounting of clinical service provision.

These needs are not unique to clinical systems, but are often neglected in business applications. In addition, healthcare systems have an underlying conflict between audit and privacy, which is not well addressed by existing systems.

Architecture. We augment messages with meta-data to allow sophisticated analysis of both local and distributed component interactions. This message meta-data provides an important bridge between application processes and the underlying middleware infrastructure. (For generality, we assume topic- or content-based publish/subscribe messaging, which also covers local or remote procedure calls.)

In our model, ordinary messages are annotated with *token* meta-data, to track information flow. Each component must send a special notification for each token it receives, describing its destination. Extra notifications are also created automatically on receipt or publication of a tokenised message. Tokens are either assigned automatically by the pub/sub system, or correspond to workflow instances, and messages are assumed to have unique identities.

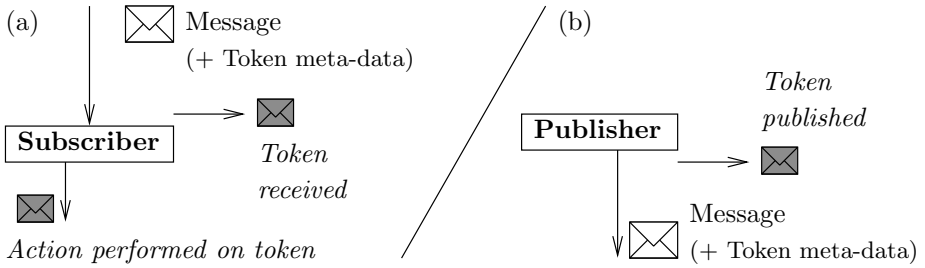


Fig. 1. The message contract for subscribers and publishers of token meta-data

Fig. 1 illustrates this contract; (a) shows a subscriber receive a message with a token, and fulfill its contract by sending notification of how it was processed. The infrastructure also sends an auxiliary message confirming token receipt. White envelopes show messages with token meta-data; grey envelopes are token notification messages. In (b), a publisher emits a new tokenised message, and the corresponding token notification message is sent automatically.

These individual *TokenAction* messages can be resequenced by matching token and message identities, to reconstruct the entire workflow as a Petri Net. This enables richer workflow reconstruction than conventional process audit logs [1], even in widely distributed systems with concurrent process execution. It also facilitates workflow cooperation, by enabling integration between messaging middleware and workflow management systems [2], and strategic software evolution.

Workflow token logging respects patient privacy, allowing rich monitoring without unnecessarily revealing patient data, based only on token logs. Tokens can also coexist with privacy meta-data, that can be enforced by pub/sub system extensions. Finally, tokens can be embedded into XML data, or treated as linked elements in their own right, e.g. in the healthcare XML schema HL7 version 3 [3].

Conclusion. This paper shows how messaging middleware and workflow systems can be extended to support the conflicting goals of reliable audit and privacy protection, especially in healthcare applications, by augmenting messages with tokens and other meta-data. This enables more precise workflow reconstruction than existing techniques based on process logs, especially in distributed applications. Monitoring these tokens enables computer systems to fulfil the technical and clinical requirements of healthcare systems, while supporting large-scale distributed operation through clean integration with secure pub/sub messaging.

References

1. Golani, M., Pinter, S.S.: Generating a process model from a process audit log. In: Business Process Management. (2003) 136–151
2. van der Aalst, W.M.P., van Dongen, B.F., et al.: Workflow mining: A survey of issues and approaches. *Data Knowl. Eng.* **47**(2) (2003) 237–267
3. Health Level Seven: HL7 reference information model (2006) [Online]. http://www.hl7.org/Library/data-model/RIM/modelpage_mem.htm [30 May 2006].

A Probabilistic Approach to Reduce the Number of Deadline Violations and the Tardiness of Workflows

Johann Eder¹, Hannes Eichner², and Horst Pichler²

¹ University of Vienna, Austria

² University of Klagenfurt, Austria

Abstract. Process prioritization strategies, based on a probabilistic temporal model, are applied to reduce the number of deadline violations and the tardiness of workflows.

Process prioritization techniques are applied to optimize process criteria, like may be the number of deadline violations or the tardiness (amount of lateness). [1,5] already showed that deadline-oriented strategies to sort work-lists, like *Earliest Deadline First*, are superior to *FIFO* or mere random selection strategies. Nevertheless, these approaches do barely consider uncertainties, which arise during process execution. They stem mainly from two aspects, undeterminable in advance: the actual duration of a task, and decisions made at conditional split points. This renders the calculation of exact temporal models at build time impossible; estimations, in the form of average values or interval representations, must be applied. To incorporate these uncertainties in a build time-calculated temporal model, we introduced a probabilistic approach. It aims at calculating temporal properties, like valid execution intervals, for each activity in the process. Temporal properties are not represented as scalars or intervals, but as *time histograms*, which are used to assess the current temporal status in a probabilistic way, e.g. to forecast the probability of a future deadline violation, or to predict the remaining execution time with a given certainty. Applied during run time, this enables pro-active features like early detection and avoidance of eventually arising deadline violations. For further details please refer to [3,2,4].

Based on this probabilistic model we introduce two new prioritization techniques. *Most Probable Deadline Violation (MPDV)* sorts the work list of a workflow participant according to deadline violation probabilities. The more likely a future deadline violation is the higher it will be ranked. This strategy aims at keeping the number of instances which violate their deadline as low as possible. *Lowest Proportional Slack (LPS)* sorts the work list according to available buffer time, and aims at minimizing the tardiness of processes. Slack can (basically) be consumed without risking a future deadline violation. Proportional slack sets the available slack in relation to the rest execution time. LPS needs a probability as input parameter – for the subsequent scenario LPS-70 (70%) proved to be ideal.

Figure 1 shows a workflow with a parallel block and two conditional blocks, augmented with branching probabilities for conditional splits. Two groups of

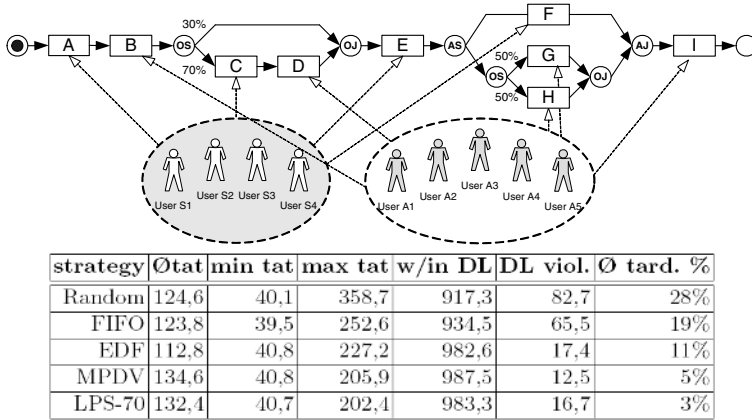


Fig. 1. Workflow Model with Resources and Simulation Results

participants are specified, along with allowed activity assignments. From the workflow log, produced by prior simulation runs, we extracted probabilistic information (branching and durations). It was used to calculate the probabilistic temporal model, which we applied at subsequent simulation runs. To generate non-uniform workload peaks, we introduced artificial burst followed by pauses. After each block of 25 started processes, with an inter-arrival frequency specified by an exponential distribution with mean 8, we inserted a random break, specified as exponential distribution with mean 50. This scenario produces high temporary work load peaks which result in long work lists (stalling behavior). The deadline was set to 190. The average simulation results (100 runs with 1000 processes each) show the turnaround time ($\bar{O}tat$, $min\ tat$, $max\ tat$), the average number of process instances that finished within the deadline ($w/in\ DL$), the average number of process instances that violated the deadline ($DL\ viol$) and the average tardiness percentage ($\bar{O}tat$).

Strategies, that are not deadline oriented produce, as expected, the highest number of deadline violations along with a high tardiness percentage. In this respect the non-probabilistic EDF-strategy performs better, excelled by MPDV, which produces the least number of deadline violations, and LPS-70, which is better suited to reduce the tardiness percentage.

References

1. G. Baggio and J. Wainer and C. A. Ellis. Applying Scheduling Techniques to Minimize the Number of Late Jobs in Workflow Systems. *Proc. of the 2004 ACM Symposium on Applied Computing (SAC)*. ACM Press, 2004.
2. J. Eder, W. Gruber, M. Ninaus, and H. Pichler Personal Scheduling for Workflow Systems. *Proceedings of the International Conference BPM'03*, Springer Verlag (LNCS 2678), 2003.

3. J. Eder and H. Pichler. Duration Histograms for Workflow Systems *Proc. of the Conf. on Engineering Information Systems in the Internet Context*, Kluwer Academic Publishers, 2002.
4. J. Eder and H. Pichler. Probabilistic Workflow Management. Technical report, Universitt Klagenfurt, Institut fr Informatik Systeme, 2005.
5. S. Rhee, H. Bae and Y. Kim. A Dispatching Rule for Efficient Workflow. *Concurrent Engineering*, Volume 12, Sage Publications, 2004.

A Unified Model for Information Integration

Ali Kiani and Nematollaah Shiri

Dept. of Computer Science & Software Engineering
Concordia University
Montreal, Quebec, Canada
{ali.kiani, shiri}@cse.concordia.ca

Abstract. We present an abstract view of information integration based on a three dimensional (3D) space of Concepts, Data Models, and Domains. In this view, the first dimension specifies the concepts (e.g., entity sets, relations, classes, etc), the second dimension defines the data model in which a concept is represented (e.g., relational, semi-structured, object-oriented, etc), and the third dimension determines the concept domain which relative to the universe of the model, uniquely identifies the application domain. We also introduce three basic transformations, called X-transform, Y-transform, and Z-transform. The queries posed to the integrated level can be expressed on the basis of these basic queries.

1 Introduction

Information integration (II) has been the subject of numerous research in Database and Artificial Intelligence, however, a main issue not yet solved is the absence of a unified/generic model for information integration. This problem has resulted in “ad-hoc” designs and hence limited use algorithms in II. Our goal in this study is to develop a model for information integration that serves as a uniform and generic access to sources in the integrated framework.

Our approach is based on defining/using some basic transformations such that query processing in a heterogeneous environment becomes (reduces to) query processing in a single/homogeneous information source.

In our model, we exploit the idea of “Generic Model Management” [1,2], for handling metadata related problems such as “schema matching”. Next, we introduce the *Conceptual Space* and the three *basic transformations* in our model.

2 Unified Model, Conceptual Space, and Transformations

In standard relational data model, queries and query processing are based on some assumptions on data model, schema of relations, unique names, and so on.

For our purpose and in our model, we try to relax some of these restrictions including unique data model, unique name, and unique schema assumptions. This would support query expression and processing, in a natural way, in the presence of heterogeneous information sources in integration.

Fig. 1 illustrates our proposed 3D view in which the X , Y , and Z axes represent concepts, models, and (application) domain, respectively. A point (C_i, M_j, S_k) in this space indicates that C_i is a concept represented in data model M_j in domain S_k . We refer to the collection of these points as “Conceptual Space.” The conceptual space corresponds to the *conceptual level* in conventional relational databases where there is no Z axis (there is only one domain) and there is no Y axis (data model is relational).

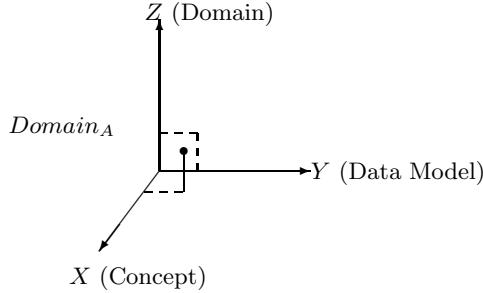


Fig. 1. Conceptual Space in the Proposed Model

In order to support query processing, we define three basic transformations, X -transform, Y -transform, and Z -transform. Given a query we try to break it down into a set of such basic transformations defined as follows.

An X -transform is a query over a domain S . In other words, an X -transform corresponds to queries over the concepts of the same domain that have the same data model. While a Y -transform converts data in a data model M_1 to data model M_2 , a Z -transform converts data in the form of a source schema to the form of a target schema.

Applying Z -transforms on the set of concepts in a query in our conceptual space unifies them into one integrated domain. Then applying Y -transforms would transform concepts in different data models into a unified data model. As a result, query processing becomes that of a single information system (i.e. X -transform).

As for future work, we are studying among other important issues in integration, constraints at the integrated level, and also properties of different data models in order to optimize the basic transformations.

References

1. P. Bernstein. Generic model management: A database infrastructure for schema manipulation. In *Proc. of Int'l Conf. on Cooperative Information Systems*. Springer-Verlag, LNCS-2172, 2001.
2. Sergey Melnik, Erhard Rahm, and Philip A. Bernstein. Rondo: a programming platform for generic model management. In *Proc. of the 2003 ACM SIGMOD Int'l Conf. on Management of data*, pages 193–204. ACM Press, 2003.

Service Composition and Deployment for a Smart Items Infrastructure

Holger Ziekow, Artin Avanes, and Christof Bornhövd

SAP Research Center LLC, 3475 Deer Creek Road,
Palo Alto, CA 94304, USA

{holger.ziekow, artin.avanes, christof.bornhoevd}@sap.com

Abstract. Smart items technologies, like RFID and sensor networks, are the next big step in business process automation. We are currently developing a dedicated Smart Items Infrastructure (SII) that supports the development and deployment of distributed ubiquitous computing applications. Decomposing these applications into individual services and deploying them at different devices poses new technical challenges. In particular, we address service composition for heterogeneous device landscapes and mapping of service implementations to devices for execution.

1 Motivation

We are currently developing a dedicated Smart Items Infrastructure (SII) that supports the development and deployment of possibly highly distributed ubiquitous computing applications and enables embedding them into business applications [1]. Ubiquitous computing applications run in the form of cooperating components or services on a variety of possibly heterogeneous devices to exploit their computing capabilities. Examples for such devices are RFID tags, sensor nodes and embedded systems.

Our goal is to bridge the gap between high level enterprise services and low level services that run on smart devices. In particular, services on smart items must be combined and orchestrated to realize the desired enterprise services. However, in the ubiquitous computing domain, decomposing applications into individual services and deploying them at different devices poses new technical challenges. One is that various different hardware platforms exist and resource limitation makes hardware abstraction usually infeasible. Hence, a composition model is required that can reflect heterogeneities on several system levels; including lower communication layers. Another challenge is to efficiently identify suitable devices for a given service in a resource-limited and dynamic system landscape. These functions are supported in the *Smart Items Infrastructure* (SII) we purpose.

2 Service Composition and Deployment with the SII

The SII contains modules to process data from smart items and for making them available to external business applications like Supply Chain Management, Asset

Tracking or Customer Relationship Management. Furthermore, modules for developing and maintaining smart items applications are provided. We divide the SII into five conceptual layers as shown in figure 1. A more detailed discussion of the components as well as the underlying system requirements and architectural decisions can be found in [1].

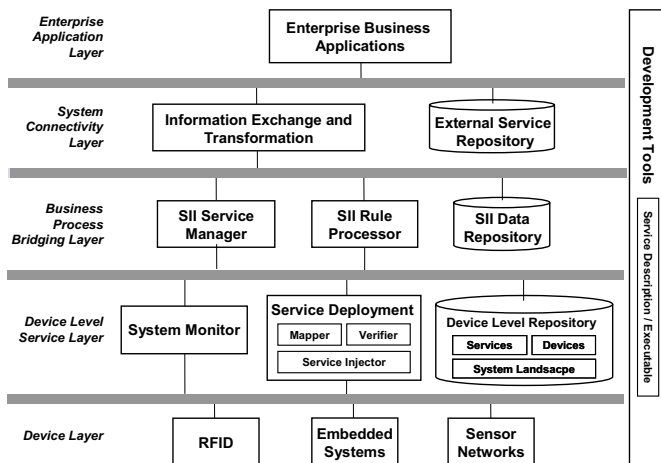


Fig. 1. Smart Items Infrastructure

The SII supports a process to compose, verify and deploy smart items applications. Therefore we developed a composition model that is based on a generic top-level ontology for describing communication issues. Service compositions are described using a derived ontology that covers domain-specific communication issues. Concepts and relations of the top-level ontology are used for generic rules that define well-formedness, compatibility and application constraints. We implemented a verification module that checks models of composite applications against these rules.

Furthermore, we implemented a mapping module to find appropriate devices for the service deployment. The module is based on a three tier architecture that structures the system regarding application domains and device characteristics. This allows for an efficient and scalable retrieval of devices profiles that fulfill the requirements of a given service. The mapping procedure considers application specific constraints and quality of service parameters. It is also used to remap services if constraints are violated due to changes in the system landscape.

Reference

1. P. Spieß, C. Bornhövd, S. Haller, T. Lin, and J. Schaper. Going Beyond Auto-Id: A Service-oriented Smart Items Infrastructure. to be published in Journal of Enterprise Information Management, March 2007.

Synthetizing RDF Query Answers from Multiple Documents

Adrian Tanasescu and Mohand-Said Hacid

LIRIS - Université Claude Bernard Lyon 1, 43 Bld. du 11 Novembre 1918
69422 VILLEURBANNE Cedex – France

{Adrian.Tanasescu, Mohand-Said.Hacid@liris.cnrs.fr}

Abstract. RDF is a recommended standard for describing knowledge about resources over the Web. If we talk about querying RDF, we must consider important aspects concerning Web querying: distributed information and context oriented description of resources. In this paper we propose a framework that provides better, more complete answering to RDF queries than classical answering mechanisms. This framework provides a way to combine several RDF documents in order to compute a more complete answer to a given query. The combination of RDF documents is performed under some conditions, leading to a safe combination.

1 Introduction

In a querying process, users search information about some resources without knowing if one document alone can satisfy the query terms. Very often, several RDF documents can contribute together to deliver an exact or approximate answer. In this paper, we investigate an approach for answering queries using several RDF documents when a single document can not provide a complete and simple answer. Combining RDF documents brings out the issue of a possible contradiction between them. In this paper, we define a measure that calculates a similarity between two RDF documents and we present an algorithm for combining several documents into a more complete answer to a given query.

2 Global Approach

2.1 Preliminary Steps

Before combining RDF documents, we introduce a preliminary phase that allows to generate a contradiction matrix capturing the dissimilarity between documents. As a second preliminary phase, we compute an abstraction tree that will avoid processing the query on the entire RDF documents database.

a. *Similarity of RDF documents*

Before query evaluation we build a similarity matrix that denotes which documents are compatible/combinable. To compute this matrix we use the following similarity measure:

$$Sim(D_i, D_j) = \begin{cases} 0 & \text{if } N_{(s)}D_iD_j = 0 \\ 1 & \text{if } N_{(s,p)}D_iD_j = 0 \\ \frac{N_{(s,p,o)}D_iD_j}{N_{(s,p)}D_iD_j} & \text{otherwise} \end{cases}$$

where $N_{(s,p,o)}D_iD_j$ is the number of triples that are identical or *approximated* as equivalent in D_i and D_j , and $N_{(s,p)}D_iD_j$ is the number of couples subject-predicate that are identical or *approximated* as equivalent in D_i and D_j .

b. Abstraction tree for query optimization

The tree is built by intersecting RDF documents into virtual nodes until no further intersection is possible. Therefore, if a virtual node can answer a query triple than all RDF documents that contributed in its computation will answer too.

2.2 Building Query Answers

Once the query is formulated, our approach follows several steps in order to compute synthetic answers:

1. *Decomposing the query into atomic goals.* The formulated query is actually decomposed in triples that will be treated separately.
2. *Generating a document-goal matrix.* This matrix contains boolean values indicating which document provides an answer to which goal. For the computation of this matrix we use the abstraction tree.
3. *Ordering candidate answers.* In order to choose the best partial answers that are to be enriched with triples from other RDF documents, we use the document-goal matrix previously calculated and order documents by the number of goals they answer.
4. *Combining RDF documents.* We choose a document D_b from the ordered list of answers. Using the similarity matrix, we retain the documents that do not contradict with D_b and order them by the number of goals (not answered by D_b) they answer. Triples from a chosen document will enrich D_b only if their subject exists in (or is subsumed by a resource in) D_b . This step is repeated for each document to be enriched.

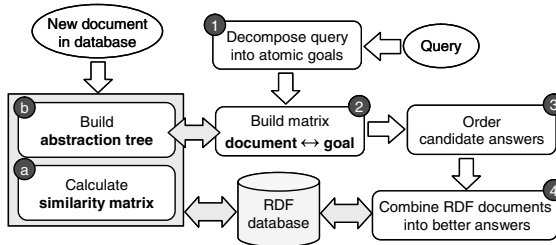


Fig. 1. Phases of the query answering algorithm

3 Conclusion

In this paper we presented an approach aimed at providing better, more complete answers to queries over RDF documents by synthetizing several RDF documents. We are currently working on the implementation of the algorithm.

Establishing Agile Partnerships in Open Environments: Extended Abstract

I.D. Stalker, M. Carpenter, and N.D. Mehandjiev

School of Informatics, University of Manchester, PO Box 88, Manchester, M60 1QD, UK
iain.stalker@manchester.ac.uk

The increasingly complex and volatile nature of many contemporary markets means that demands are often best satisfied through dynamic networks of collaborating enterprises. Successful collaboration demands tight, flexible integration of business processes, however, this assumes that an appropriate team has been assembled. Traditionally, a top-level service or goal is decomposed into component services or subgoals each of which is then matched to a provider. This is a complex task and while automated tools exist, supported especially by the notions of service discovery and traders, significant guidance is typically sought from the user. This imposes a substantial burden of interaction and considerable knowledge is demanded of a user to decompose to a level of detail which allows for matching to known services. Problems arise if this is not the case and open environments, such as the internet, present additional difficulties: if a user is not up-to-date potential decompositions may be missed; new entrants into a market may not be recognised; etc. Bottom-up approaches circumvent some of these difficulties, but also come at a price. For example, where goal decompositions are available, these are typically much more inefficient; if there is only a fixed number of processes available within a system, the case of “no solution” may take considerable time to establish. Moreover, since many bottom-up approaches distribute control, the system is vulnerable to malicious behaviour. Thus, a certain level of trust is required.

We present a novel synthesis of top-down and bottom-up techniques for cross-organisational team formation which leverages the benefits of each while minimising their disadvantages. The work developed in the context of the EC-Funded Project CrossWork (www.crosswork.info), which focuses upon distributed, cross-organisational development processes in the automotive industry. We target so-called *Networks of Automotive Excellence (NoAEs)*. An NoAE provides a environment of automotive service providers and suppliers in which appropriate agile partnerships (*teams*) can emerge to respond to a market opportunity as a single virtual entity. In this way (members; subgroups of) the NoAE can compete with established Tier 1 suppliers for orders from *Original Equipment Manufacturers (OEMs)*, thus promoting a sustained growth in one of the most competitive business sectors. The dynamic nature of the underlying NoAE makes it a challenge to develop software systems which automate the creation of such teams and cross-organisational workflow models. Indeed, this is considered an open research issue. In CrossWork, we have created an enabling infrastructure to realise (semi-)automatic team formation, (local) workflow formation and configuration, and enactment of the resulting global workflow within NoAEs.

A team forms with the delivery of some specific product in mind: this is its goal. The function of the team is to transform the current state, characterised by possibly incomplete design information, a statement of raw materials, etc. into an attainment

of this goal. Initially, this transformation is too abstract to match with actual services, available expertise, and so forth. Thus, it needs to be elucidated. Information about the physical product is used to identify combinations of subgoals which can be matched to (known) services. These provide abstract processes through which the global goal can be achieved. Team formation takes an abstract process, finds particular providers for each service, identifies supporting services, e.g. logistics, and refines it into a concrete process. Its primary activity is intelligent matchmaking.

The top-down approach assumes access to appropriate syntactic and semantic structures to allow goals to be rewritten to a level of detail which facilitates matching to service providers. Naturally, this assumes knowledge of service providers or at least some indirect access to these. An additional disadvantage is the possible need to make several approaches to a number potential service providers owing to lack of information about workloads and availability (cf. opportunism). To overcome some of these difficulties, the top-down approach is complemented by a bottom-up approach which is inspired by *Blackboard Systems (BBSs)*, e.g. [3]. A BBS formalises the metaphor of a group of experts communicating ideas using a blackboard. The flow of information between blackboard and expert is bi-directional. In the bottom-up approach, a “noticeboard” is created which summarises a business opportunity, including *inter alia* initial state and desired final state. Interested and available parties contribute suggestions, leading to intermediate states which develop the initial state or the final state; and which themselves can be developed by others. It is possible that a state may only be partly developed, giving rise to branching. From these interactions a “global solution” is found and the appropriate team emerges. A useful benefit is that the approach allows for a provider to customise its processes, for example, if a particular opportunity is lucrative. A fuller discussion of the approach can be found in [1] and [2].

For well-defined products, a dedicated top-down approach is usually appropriate. Where novel products are being developed, a bottom-up approach is often the only option owing to a complete lack of semantics. There are occasions where a combined approach represents an exciting possibility. Consider, for concept design and early stages of product development, even when a functional decomposition is possible, bottom-up approaches can significantly enhance creativity and foster synergies which may be otherwise missed. Moreover, where some suppliers provide more than one part of an automotive system, a detailed decomposition can be redundant; and typically the exact decomposition of the system is negotiated by the team members. To minimise the shortcomings of each approach and maximise the benefits of each, we typically use a top-down approach to obtain a high-level decomposition; each resulting subgoal is used to create a noticeboard to drive a bottom-up approach through which the remainder of the solution can emerge.

References

1. Martin Carpenter, Nikolay Mehandjiev, and Iain Stalker. Emergent process interoperability within virtual organisations. In *Proceedings of ATOP 2005, at AAMAS 2005*, 2005.
2. Martin Carpenter, Nikolay Mehandjiev, and Iain Stalker. Flexible behaviours for emergent process interoperability. In *Proceedings of PINCET 2006, at WETICE 2006*, 2006.
3. I Craig. *Blackboard Systems*. Ablex, Norwood, NJ, 1995.

Distributed Adaptation Reasoning for a Mobility and Adaptation Enabling Middleware

Nearchos Paspallis and George A. Papadopoulos

Department of Computer Science, University of Cyprus,
75 Kallipoleos Street, P.O. Box 20537, CY-1678, Nicosia, Cyprus
{nearchos, george}@cs.ucy.ac.cy

Abstract. The prospect of adaptive, mobile applications provides both opportunity and challenge to the application developers. Adaptive, mobile applications are designed to constantly adapt to the contextual conditions with the aim of optimizing the quality of their offered service. In this respect the MADAM project provides software engineers with reusable models, tools and runtime support for enabling adaptive behavior in their mobile applications. This paper presents an extension to the MADAM middleware architecture which enables distributed compositions. To this end, a new adaptation reasoning approach is described, which improves on the original one in two ways: it allows decentralized reasoning for selecting the most suitable adaptation and it supports distributed application composition. Moreover, the proposed approach is argued to provide additional benefits such as robustness, agility and scalability.

Adaptive, mobile and pervasive computing applications are designed to constantly adapt to their contextual conditions in an autonomous manner. The aim of the adaptation is to optimize the quality of the service offered to the end users. This study builds on the results established by existing solutions [1, 2], and extends them to introduce mechanisms for enabling distributed adaptation reasoning while at the same time maintaining attributes such as robustness, agility and scalability.

Modern approaches define adaptive, component-based applications as collections of software components which can be configured according to a number of different compositions. Furthermore, technologies such as reflection and component orientation enable reasoning and possibly altering of their behavior. An important question though is *how* can the underlying middleware automatically reason about the context and *select* an optimal composition to adapt to. This question becomes further challenging as additional attributes such as robustness, agility and scalability are aimed.

The proposed approach depends on composition plans which are defined at design time and which can be used to dynamically construct different variants of the application. Individual variants are designed so that they offer certain advantages, such as for example better resource utilization for a particular context. Naturally, each variant is designed with the aim of maximizing the utility of the application for at least some points in the context space. While multiple options are possible for the realization of the adaptation reasoning, utility functions offer the important advantage of supporting adaptation reasoning of components which become available after deployment.

Utility functions can be used to compute the *utility*, i.e. a quantifiable measure of the quality of the service as it is experienced by the application users. In this respect, the overall objective of the middleware can be defined as the continuous evaluation, and selection of a composition which maximizes the utility. Any knowledgeable decision requires that the reasoning process is aware of the contextual information of all the parts involved. In distributed environments, this implies that the contextual information of all participating hosts must be communicated to the host which performs the adaptation reasoning.

The continuous computation of the utility values can be quite costly though, especially in frequently changing environments such as in mobile and pervasive computing settings. In this respect, two custom-tailored adaptation reasoning approaches are introduced: *proactive* and *reactive* adaptation reasoning. These two approaches differ in the timing of the adaptation reasoning and its required steps. Proactive reasoning requires that all context data is communicated as soon as it becomes available. Contrary to this, reactive adaptation reasoning defers the communication of such context data until they are actually needed. Additional (hybrid) approaches are also possible.

Both options provide individual benefits which make them better choices, depending on the particular requirements of the application. For example, the proactive approach is more likely to achieve faster and more accurate decisions as more context data is available to the decision making process at any moment. In contrast, the reactive approach is better in terms of resource consumption as the context data are communicated only when needed. The latter benefit becomes more important when the context changes more often than the rate at which the application is needed to adapt.

Besides the benefits of the two individual adaptation reasoning approaches, it is argued that an implementation middleware can also benefit by using hybrid approaches, or by dynamically switching from one approach to the other on demand. Furthermore, it is argued that the use of utility functions in this approach not only enables *proactive* and *reactive* adaptation reasoning, but also enables the construction of distributed protocols which satisfy the *robustness*, *agility*, and *scalability* requirements. Robustness is achieved by means of supporting applications to re-compute and failover to alternative, possibly centralized, compositions when a failure (e.g. network outage) prevents the originally selected adaptation. Additionally, the ability to re-compute and implement only a part of a composition greatly improves the agility of an application, especially in the case of the proactive approach. Finally, scalability is achieved as a means of the protocol support for distributed, decentralized computation of the utility functions and construction of compositions.

References

1. Floch, J., et al., Using Architecture Models for Runtime Adaptability, Software, IEEE, 2006. Volume 23 (Number 2): p. 62-70.
2. Paspallis, N. and G.A. Papadopoulos, An Approach for Developing Adaptive, Mobile Applications with Separation of Concerns, to appear in the 30th Annual International Computer Software and Applications Conference (COMPSAC), Chicago, IL, USA, September 16-21, 2006: IEEE.

Scheduling of Composite Web Services

Dmytro Dyachuk and Ralph Deters

Department of Computer Science, University of Saskatchewan
Saskatoon, Saskatchewan, S7N 5C9 Canada
dmytro.dyachuk@usask.ca, deters@cs.usask.ca

Abstract. Composite Web Services (CWS) aggregate multiple Web Services (WS) in one logical unit to accomplish a complex task (e.g. business process). This aggregation is achieved by defining a workflow that orchestrates the underlying Web Services in a manner consistent with the desired functionality. Since CWS can aggregate atomic WS and/or other CWS they foster the development of service layers and reuse of already existing functionality. An important issue in the deployment of services is their run-time performance under various loads. Due to the complex interactions of the underlying services, a CWS they can exhibit problematic and often difficult to predict behaviours in overload situations.

This paper focuses on the use of request scheduling for improving CWS performance in overload situations. Different scheduling policies are investigated in regards to their effectiveness in helping with bulk arrivals.

Keywords: Web Services, Composite Web Service, LWKR, SJF, Scheduling, Admission Control.

Web Services are most often used to expose some already existing legacy functionality (e.g. transactional database). For a service governed by the Processor Sharing scheduling policy (PS), an unexpected increase of the request arrival rate can lead to an overload and ultimately the *thrashing effect* [1]. Thrashing occurs either as a result of an overload of physical resources (*resource contention*) like a processor or memory or as a result of locking (*data contention*). Our experiments showed that the majority of Web Services are vulnerable to the thrashing effect. Thus an attempt to handle a large bulk of requests results in a significant performance decrease. Thrashing is particularly problematic in workflows, since the throughput of the flow in the network is equal to the throughput of the “slowest” intersection. Consequently thrashing of one service will therefore negatively impact the performance of the CWS. In this paper we propose applying scheduling policies such as Shortest Job First (SJF), and Least Work Remaining (LWKR) in combination with adaptive load control [1] to Web Services. Scheduling of CWS requests can be done on a *system* (workflow/CWS) or *component* (service) level. Adding scheduling and an admission control to an already existing service (e.g. WS, CWS) can be achieved by use of the proxy pattern. The proxy buffers all the excessive requests, thus prevents service from overload. In component-level scheduling a separate scheduler is placed in front of each service resulting in a multi-step scheduling. For system level scheduling only one scheduler is needed and the scheduling is done on a CWS request admission phase.

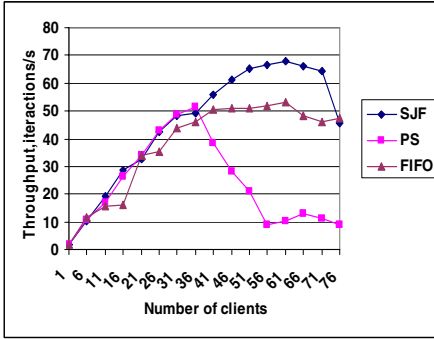


Fig. 1. Throughput of the service governed by various scheduling policies

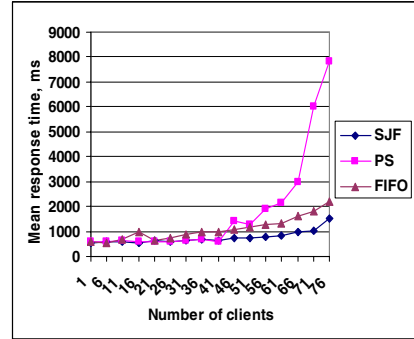


Fig. 2. Response times of the service governed by various scheduling policies

FIFO represents the use of admission control that uses a queue to buffer requests once the saturation point of the provider is reached. SJF uses a priority queue, prioritizing short jobs at the expense of larger ones. LWKR evaluates the total cost of all remaining sub-request and assigns the priorities according to the remaining work.

Using the TPC-APP [3], a two-tier B2B application, as a benchmark we evaluated the performance gains of SJF (Shortest Job First) compared to pure admission control and standard PS for a single service. By introducing a proxy between a consumer and a provider, it is possible to achieve a transparent scheduling and admission control that leads to significant performance improvements in overload cases. In addition, the experimental evaluation showed that even in the absence of a priori knowledge, a SJF scheduler that uses observed runtime behaviour can lead to schedules that outperform FIFO and PS, making it an attractive approach for boosting Web Services performance. The results of the experimentation indicate that transparent scheduling can be applied to Web Services as an effective and easy to implement approach for boosting performance and avoiding service provider thrashing. LWKR improves the performance up to 68%, but requires placing a proxy in front of each component. Using Shortest Job First (SJF) and having just one scheduler for the workflow leads to a 7% drop in performance compared to LWKR.

References

1. H.-U. Heiss, and R. Wagner, "Adaptive load control in transaction processing systems", In 17th International Conference on Very Large Databases, Barcelona, Spain, September 1991.
2. P. Brucker, Scheduling Algorithms Third Edition. Springer-Verlag, Berlin, Germany, 2001.
3. TPC-APP. http://www.tpc.org/tpc_app/
4. Workflow Patterns www.workflowpatterns.com/

Handling and Resolving Conflicts in Real Time Mobile Collaboration

Sandy Citro, Jim McGovern, and Caspar Ryan

School of Computer Science and Information Technology,
RMIT University, Melbourne, Australia
scitro@cs.rmit.edu.au, jim.mcgovern@rmit.edu.au,
caspar@cs.rmit.edu.au

1 Introduction

Real time group editors allow two or more users at different locations to work on a shared document at the same time. In a mobile network environment with non-deterministic communication latency, a replicated architecture is usually adopted for the storage of the shared document in order to provide high responsiveness. A conflict occurs when two or more users have different intentions for editing the same part of the replicated document. Conflict can be categorised into two types: *exclusive* and *non-exclusive* conflicts. An *exclusive* conflict occurs when the conflicting operations cannot be realised at the same time, and if serially executed, the effect of the later operation will override the earlier operation. In contrast, a *non-exclusive* conflict occurs when the conflicting operations can be realised at the same time and both operations can be applied to the target without one overriding the other.

The approaches adopted by the existing algorithms to resolve conflicts can be categorised into: (1) locking approaches [1], (2) operational transformation approaches [2], and (3) multi-versioning approaches [3]. The locking approach is a conflict prevention approach rather than a conflict resolution approach, and it does not promote concurrency as only one person can modify an object at one time. The operational transformation approach is optimistic and is able to handle non-exclusive conflicts, but it cannot handle exclusive conflicts. The multi-versioning approach handles exclusive conflicts by creating different object versions with each version realising each conflicting intention. Xue et al. [4] combine multi-versioning, operational transformation, and post-locking to handle both types of conflict and to restrict the number of object versions. However, while post-locking simplifies the conflict resolution process by locking versions to protect them from further modification, it suffers from a *partial intention* problem insofar as conflicts could be better resolved if the conflicting intentions have been completely realised and thus every user can see the full intention of all other users.

2 Contribution

The authors have proposed a conflict management algorithm that aims to be suitable for mobile environments. The algorithm combines operational transformation and multi-versioning to handle exclusive and non-exclusive conflict, while respecting user

intention at the semantic level. While it is beyond the scope of a short paper to describe the algorithm at a high level of technical detail, the general structure and behaviour of the algorithm is given below.

Firstly, for non-exclusive conflicts, operational transformation is used to transform one of the conflicting operations against the other to preserve both intentions. Secondly, for exclusive conflicts, the multi-versioning approach is used to preserve both intentions. Thirdly, the algorithm implements a variation of post-locking called *delayed post-locking* to address the partial-intention problem by allowing users to completely realise their intention without being interrupted by incoming conflicting operations. The delayed post-locking technique uses a lock called a *user intention lock (UIL)* to prevent any interruption from incoming operations allowing the user to fully realise his/her intention. Finally, the novel use of a *conflict table* is implemented to store all conflict information for the purpose of facilitating conflict resolution. Whenever a conflict occurs, the object in conflict is locked, the user is notified, and the operations causing the conflict are stored in a conflict table. Each entry in the Conflict Table represents the conflicting object version and each entry has an attribute to inform users whether or not the intention on the particular object version has been completely realised. This information allows users to make better conflict resolution decisions in terms of which document version should be accepted. A conflict is eventually resolved by selecting one Conflict Table entry to be the final version for that particular conflict. The proposed algorithm can be used with various conflict resolution strategies, whilst satisfying the following properties: it avoids the partial-intention problem; it need not depend on a group leader or other pre-specified conflict resolution roles; the conflict can potentially be resolved without having to wait for all sites to receive all conflicting operations (dependent upon the chosen conflict resolution strategy); and finally, the algorithm provides better information to users so that they can resolve the conflict knowing the status of the conflict.

Future work will involve integrating the proposed algorithm with consistency management, membership management and document partitioning algorithms to provide a comprehensive framework for real time mobile collaboration. Ongoing work will also look at alternative conflict resolution strategies, with particular emphasis on their effectiveness from a usability point of view, and their performance and impact on resource consumption within a mobile environment.

References

1. Munson, J., Dewan, P.: A concurrency control framework for collaborative systems. In: Proceedings of the 1996 ACM conference on Computer supported cooperative work, ACM Press (1996) 278–287
2. Ellis, C.A., Gibbs, S.J.: Concurrency control in groupware systems. In: Proceedings of the 1989 ACM SIGMOD international conference on Management of data, ACM Press (1989) 399–407
3. Sun, C., Chen, D.: Consistency maintenance in real-time collaborative graphics editing systems. ACM Transactions on Computer-Human Interaction 9(1) (2002) 1–41
4. Xue, L., Zhang, K., Sun, C.: An integrated post-locking, multi-versioning, and transformation scheme for consistency maintenance in real-time group editors. In: ISADS. (2001) 57–64

A Configurable Event Correlation Architecture for Adaptive J2EE Applications

Yan Liu^{1,2}, Ian Gorton^{1,2}, and Khanh Vinh Le³

¹National ICT Australia, NSW, Australia*
School of Computer Science and Engineering

²University of New South Wales, Australia

³Faculty of Information Technology,
University of Technology Sydney, Australia
{jenny.liu, ian.gorton}@nicta.com.au,
Vinh.Le@students.uts.edu.au

Abstract. Distributed applications that adapt as their environment changes are developed from self-managing, self-configuring and self-optimising behaviours. This requires constant monitoring of the state of the environment, and analysing multiple sources of events. Event correlation is the process of correlating monitored events from multiple sources for further analysis. It is essential that event correlation supports reliable event management with minimal delay. This paper describes the design and implementation of an event correlation architecture for adaptive J2EE applications. The architecture supports flexible configuration of event correlation in terms of the reliability and performance. This is especially useful in situations when multiple sources of events have different level of requirements for reliability and performance. We evaluate the performance overhead of this event correlation architecture and demonstrate its practical usage in a case study of an adaptive image server application.

Keywords: Event correlation, adaptive, distributed application, message, middleware.

1 Introduction

Application servers provide a standardized service layer to support the development and deployment of distributed, server-side applications. However, such systems remain challenging to construct, as they operate in changing environments with variable user loads, resource levels and unpredictable system faults. Deciding on the 'best' parameters settings is difficult, and administrators are forced to perform tuning and configuration through observation and experimentation every time an environmental change occurs. This results in on-going high administrative overheads and costs for managing the system.

The solutions inspired by autonomic computing [1] are enabling application server technologies with the ability to adapt in response to functional and

* National ICT Australia is funded through the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

environmental changes that occur after application deployment. This adaptation depends on constantly monitoring the systems and adjusting configuration parameters.

Event correlation allows monitoring and filtering of events among components [2]. Event correlation plays a crucial role in reducing network traffic and computation time. Therefore correlation engines are core infrastructure components for constructing adaptive distributed server applications with J2EE.

2 The Architecture for Configurable Event Correlation

We propose an architecture solution shown in Figure 1 to implement configurable event correlation for J2EE-based applications with adaptation. We address the interoperability of this architecture for integration with different adaptive applications by utilizing the Common Base Event format. The internal structure of the ECE allows event correlation to be integrated with different actions. A unique feature of this architecture is that it enables the configuration of QoS to tune reliability and performance at the message level. This is achieved by introducing a message queue and proxy components that transparently interact with the underlying queuing technology. Our empirical evaluation is carried out in an adaptive image server application, which scales an image size and its resolution to reduce network overhead and increase performance. The results demonstrate that this architecture is extensible and lightweight with good performance and scalability.

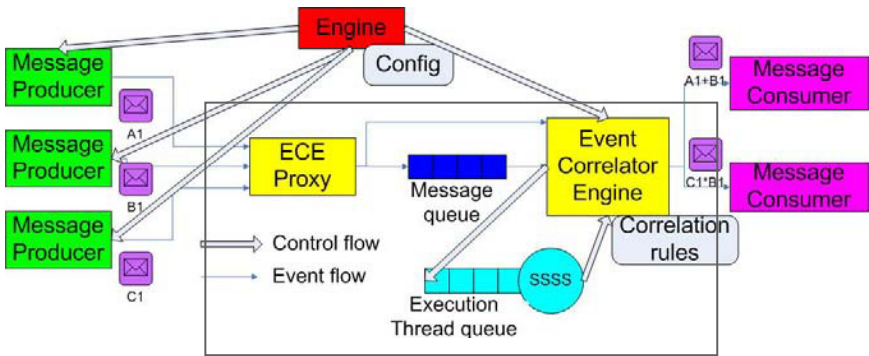


Fig. 1. The architecture of the configurable event correlation engine

The current version of ECE only supports sequence-based event correlation. Our on-going work is extending ECE with other event correlation pattern. Another limitation of this architecture is that the event correlation is not integrated with advanced knowledge-based rule and policy management mechanisms used in autonomic computing systems. With the foundation of this event correlation architecture established, we are planning to further extend its capabilities in this direction.

References

- [1] Kephart, J. O., Research challenges of autonomic computing, ICSE '05: Proc. of the 27th Intl. Conf. on software engineering, (2005), 15-22.
- [2] Stojanovic, L., Schneider, J., Maedche, A., Libischer, S., Studer, R., Lumpp, T., Abecker, A., Breiter, G., and Dinger, J. 2004. The role of ontology in autonomic computing systems. *IBM Syst. J.* 43, 3 (Jul. 2004), 598-616.

Autonomous Deployment and Reconfiguration of Component-Based Applications in Open Distributed Environments

Jérémy Dubus and Philippe Merle

INRIA Futurs, Jacquard project
Laboratoire d'Informatique Fondamentale de Lille - UMR CNRS 8022
Université des Sciences et Technologies de Lille - Cité Scientifique
59655 Villeneuve d'Ascq Cedex France
Jeremy.Dubus@inria.fr, Philippe.Merle@inria.fr

Abstract. In open distributed environments (ODEs), such as grid and ubiquitous computing, deployment domains can not be statically identified as they dynamically evolve. Thus, ADLs are unadapted to describe explicitly and exhaustively applications deployed and executed on ODEs. We argue that concepts for managing evolution autonomously should allow architects to describe how their component-based applications must evolve when the deployment domain evolves too. The contribution of this paper is DACAR, a rule-based framework to address autonomous evolution of software architectures in open distributed environments.

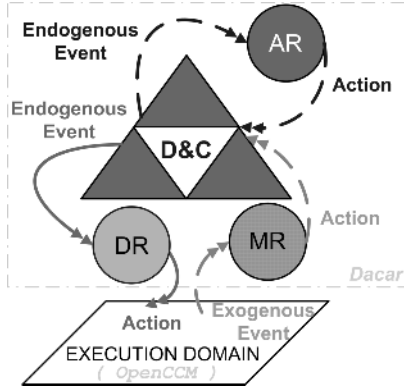
1 Problem Statement

Classically, distributed applications are deployed on computers interconnected through a local network, where the *deployment domain* —*i.e.*, the set of nodes available to host application components is statically known. Nevertheless, in ODEs, such as grid or ubiquitous computing, the list of available nodes evolves dynamically. Distributed systems are built using the concept of architecture of software components. These components are associated with the deployment domain. These assemblies are described using an *Architecture Description Language* (ADL). Using ADLs to design applications in ODEs seems to be unadapted. Associating software component instances with nodes to deploy a distributed application in ODEs is impossible since the deployment domain is unknown at design-time. Furthermore, at runtime, nodes that are hosting a part of the application components can enter in or leave the deployment domain.

Our contribution is DACAR a framework that allows the designer to define the abstract architecture (using an ADL), and to add rules to describe architectural evolutions. These rules define the *adaptation policy* —*i.e.*, the set of desired reconfigurations for the architecture depending on events likely to happen. The description and its relative adaptation policy are then transformed into a running component-based autonomous system. This approach is inspired by the *autonomic computing* paradigm [2].

2 Contribution

The global architecture of DACAR represented on the figure is separated in three parts. The *reified architecture* (“D&C” on the figure) is a dynamically modifiable model based on OMG D&C [3]. This model is equivalent to an ADL description.



The *execution domain* encompasses the runtime artifacts of the application as well as the deployment domain. This domain is dynamically reconfigurable in order to deploy new runtime artifacts.

Finally, the *reconfiguration rules* are *Event-Condition-Action* (ECA) rules [1]. They are triggered by events coming either from the reified architecture or the execution domain. The rules are responsible of ensuring the causal link between the reified architecture and the execution domain. They are also responsible of the

high-level business self-adaptation policy of the application. There are two types of events: *endogenous* events are sent from the reified architecture, and *exogenous* ones are sent from the execution domain. Reconfiguration rules are classified into three categories. **Deployment Rules (DR)** are triggered by an endogenous event. They operate an execution domain-level reconfiguration of the concrete architecture. DRs ensure that every relevant change in the reified architecture is executed on the execution domain. **Monitoring Rules (MR)** are triggered by an exogenous event. They operate a reified architecture-level reconfiguration. MR aim to update the reified architecture in order to notify changes at the execution domain level. Both MR and DR are independent of the application. **Architectural Rules (AR)** are triggered by an endogenous event. They operate a reified architecture-level reconfiguration. They provide a way to ensure architectural properties dependant upon changes from the architecture itself. AR are specific to the application. They represent the adaptation policy of the application to be written by the administrator of the application.

DACAR is implemented with the Fractal component model and tested to deploy autonomous CORBA component-based applications from OMG D&C descriptions on top of the OpenCCM platform.

References

1. Thierry Coupaye and Christine Collet. Denotational Semantics for an Active Rule Execution Model. *2nd International Workshop on Rules in Database Systems, LNCS*, pages 36–50, 1995. London, UK.
2. Jeffrey Kephart and David Chess. The Vision of Autonomic Computing. Technical report, IBM Thomas J. Watson, January 2003. IEEE Computer Society.
3. OMG. Deployment and Configuration of Distributed Component-based Applications Specification. Version 4.0 formal/06-04-02, April 2006.

Dynamic Integration of Peer-to-Peer Services into a CORBA-Compliant Middleware

Rüdiger Kapitzka¹, Udo Bartlang², Holger Schmidt³, and Franz J. Hauck³

¹ Dept. of Comp. Sciences, Informatik 4, University of Erlangen-Nürnberg, Germany
rrkapitz@cs.fau.de

² Siemens AG, Corporate Technology, Munich, Germany
udo.bartlang.ext@siemens.com

³ Distributed Systems Laboratory, University of Ulm, Germany
{holger.schmidt, franz.hauck}@uni-ulm.de

Abstract. Peer-to-Peer computing has evolved over the last few years and is applied to a rising number of applications. Following this development we present a decentralised approach to dynamically select, load and integrate peer-to-peer based services into a CORBA-compliant middleware. This is achieved by extending and improving the mechanisms for dynamic service integration of JXTA an open peer-to-peer infrastructure. At object level we build on the fragmented object model provided by the AspectIX middleware to seamlessly integrate and use peer-to-peer services instead of common CORBA client/server-based implementations.

Common object-oriented middleware infrastructures provide mature support for client/server programming but usually miss facilities to easily and seamlessly support other interaction paradigms. We present a flexible approach to integrate peer-to-peer based services into a CORBA-compliant middleware at object level based on our AspectIX middleware. This offers the possibility to replace client/server interaction by peer-to-peer communication without any client-side modification.

The integration of peer-to-peer services at object level is achieved by a modularisation of the object reference (IOR) handling using a *generic reference manager* with *portable profile managers* [2]. These profile managers encapsulate all tasks related to reference handling, i.e., reference creation, reference marshalling and unmarshalling, and type casting of representatives of remote objects. Currently, AspectIX provides profile managers for standard CORBA and additionally offers support for the fragmented object model [4] and other non-CORBA middleware platforms, like Jini or Java RMI. In this work, we implemented a profile manager and specified an own JXTA IOR profile for integrating the JXTA middleware [1] and its peer-to-peer services based on the support for the fragmented object model.

A fragmented object might be composed of several fragments and could be distributed over multiple machines. While a fragmented object offers a standard object interface to the outside, it can exhibit an arbitrary architecture inside. The fragmented object model provides maximal flexibility but might result in a higher development effort especially for recurring demands like fault-tolerance or a peer-to-peer-based internal communication. The integration of JXTA to support peer-to-peer interaction eliminates this drawback as it comprises a fully-fledged middleware.

However, the JXTA middleware cannot be compared to a standard object-based middleware supporting the client/server paradigm. Instead, it misses the concept of an object reference and an application-directed interface, but offers lower level network-directed concepts like a service description specifying a network protocol and an abstract communication endpoint to connect other services supporting the same protocol. Thus, our JXTA IOR profile contains a metadata description called *module specification advertisement* describing a JXTA service and its supported protocol. A fragmented object can now internally consist of JXTA service instances interacting in peer-to-peer fashion. By combining the fragmented object model and the JXTA service concept we bridge the gap between a standard object-based client/server middleware and the JXTA peer-to-peer infrastructure.

When coping with a fragmented object, the binding usually requires the dynamic loading of code, as it is not feasible to install and load all code modules at every node of the system. The reason is that these would only be used by some nodes and these may even not be known in advance. Therefore, we recently proposed a dynamic loading service that enables the dynamic loading of platform-specific code on demand [3]. In contrary to that work, we propose a generic and decentralised peer-to-peer-based lookup, selection and loading process. This allows multiple parties to independently provide implementations for a certain object. The current prototype extends existing concepts of the JXTA platform to dynamically select and load code, based on advertisements and extends those to provide a truly platform-independent support for the dynamic loading of platform-specific code. Our architecture is composed by three components: A *decentralised implementation repository* represented by a JXTA peer group hosting metadata about all available objects and their implementations. A *code provider*, which is run by every entity offering code, that is responsible for publishing implementation-related advertisements via the implementation repository and sharing the code. Finally, there is the *dynamic loader*, that builds the core of our prototype. It is invoked by the JXTA profile manager during binding time of a JXTA-based fragmented object to select and load a platform specific implementation.

Even though our current prototype enables the dynamic integration of peer-to-peer-based services by a precise selection of platform-specific code, we currently assume that an implementation is self-contained. We intent to investigate solutions to support implementations that reference other implementations enabling more modularity.

References

1. Li Gong. JXTA: A network programming environment. *IEEE Internet Computing*, 5(3):88–95, 2001.
2. Franz J. Hauck, Rüdiger Kapitza, Hans P. Reiser, and Andreas I. Schmied. A flexible and extensible object middleware: CORBA and beyond. In *Proc. of the Fifth Int. Workshop on Software Engineering and Middleware*. ACM Digital Library, 2005.
3. R. Kapitza and F.J. Hauck. DLS: a CORBA service for dynamic loading of code. In *Proc. of the OTM Confederated Int. Conf.*, Sicily, Italy, 2003.
4. Mesaac Makpangou, Yvon Gourhant, Jean-Pierre Le Narzul, and Marc Shapiro. Fragmented objects for distributed abstractions. In T. L. Casavant and M. Singhal, editors, *Readings in distributed computing systems*, pages 170–186. IEEE Computer Society Press, 1994.

Automated Deployment of Enterprise Systems in Large-Scale Environments

Takoua Abdellatif¹, Didier Hoareau², and Yves Mahéo²

¹ BULL SA, LSR-IMAG / INRIA

Takoua.Abdellatif@inrialpes.fr

² Valoria Lab – University of South Brittany

{Didier.Hoareau, Yves.Maheo}@univ-ubs.fr

Abstract. The deployment of multi-tiered applications in large-scale environments remains a difficult task: the architecture of these applications is complex and the target environment is heterogeneous, open and dynamic. In this paper, we show how the component-based approach simplifies the design, the deployment and the reconfiguration of a J2EE system. We propose an architecture description language that allows specifying constraints on the resources needed by the components and on their location, and a deployment solution that handles failures.

Introduction. J2EE application servers are complex service-oriented architectures. They are generally deployed on clusters to improve their quality of service. A J2EE cluster is composed of replicated Web and EJB tiers for load balancing and fault tolerance. A front-end load balancer dispatches the HTTP requests to the containers. In large-scale environments, machines are highly distributed and heterogeneous in terms of software and hardware configurations. Furthermore, these resources can be dynamic. Therefore the resource allocation should be automated and the deployment process should automatically take into account the dynamicity of the environment. We make the assumption that the large-scale environment is structured in zones and that for each zone are defined some known machines called *zone managers* whose role is to maintain a list of the machines in the zone and to orchestrate the deployment process.

Deployment System. We adopt an architecture-based approach to manage the J2EE system. We wrap system parts into explicitly bound components. The obtained system, called *JonasALaCarte*, is based on the Fractal component model.

In large-scale environments, we cannot know in advance the target machine for each component of the system. So, in order to specify the deployment of a J2EE system, we have added to the Fractal architecture descriptor (that defines the architecture of the system in terms of component definitions and component bindings) a *deployment descriptor*, that contains, for each component, the description of the resources that the target platform must satisfy, and references to component instances. The deployment descriptor lists all the constraints that a hosting machine has to verify. *Resource constraints* allow hardware and software needs to be represented, and *location constraints* make it possible to control the placement of a component when more than one host apply for its hosting. These constraints are solved thanks to our deployment process that allows, additionally, the recovery from failures.

Deployment Process. For each zone in the environment, a zone manager maintains the list of the machines in the zone that may host the J2EE system components. This zone manager is given the architecture and deployment descriptors by an administrator. It then multicasts them to the zone nodes. Each node checks the compatibility of its local resources with the resources required for each component. If it satisfies all the resource constraints associated with a component, it sends to the manager its candidature for the instantiation of this component. The manager receives several candidatures and tries to compute a placement solution in function of the location constraints and the candidatures. The manager updates the deployment descriptor with the new placement information and broadcasts it to all the zone nodes. Each node that receives the new deployment descriptor updates its own one and is thus informed of which components it is authorized to instantiate and of the new location of the other components. The final step consists in downloading necessary packages from well defined package repositories whose location is defined in the deployment descriptor.

The steps described above define a *propagative deployment*, that is, necessary components for running J2EE applications can be instantiated and started without waiting for the deployment of all the components. As soon as a resource become available or a machine offering new resources enters the network, candidatures for the installation of the “not yet installed” components will make the deployment progress.

Some preliminary experiments we have conducted on a prototype implementation show that the performance of the resource observation and the constraint solving remain acceptable even for a large number of non trivial resource constraints.

Automatic Recovery from Failures. In the environment we target, resources can also become unavailable, some parts of the J2EE system can be faulty and some machine may fail. In this work, we consider silent failures. When a component does not respond to a method call or a request within a timeout, the node detecting the failure sends to the zone manager a message holding the identity of the component to redeploy. Then, the zone manager updates the deployment descriptor by removing the location of the component and broadcasts the new descriptor to all the machines connected in the zone. This automates the redeployment of the faulty component since all the machines find themselves back in the propagative deployment described above.

Specific actions are carried out in the case of the failure of replicated components (eg EJB container and Web container services) or zone managers, exploiting group communication and temporary storage of incoming requests.

Conclusion. The work described in this paper proposes a solution for the deployment of enterprise systems in large-scale environments. Our main contribution consists in the following points. First, the deployment system is resource-aware and the constraint resolution is performed in a reasonable time. Second, the deployment task is simplified since the administrator role is reduced to writing a deployment descriptor. All the deployment process and the recovery from failures are automated. Finally, we maintain the performability of the system since we maintain the structure described in the architecture descriptor by replacing each time a faulty component by another. This allows assuring the continuity of Internet services and maintaining their quality of service.

Supporting Reconfigurable Object Distribution for Customizable Web Applications

Po-Hao Chang and Gul Agha

University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
{pchang2, agha}@cs.uiuc.edu

1 Introduction

Web applications are tightly coupled with the platforms that a particular service provider intends to support and the execution scenario envisioned at the design time. Consequently, the resulting applications do not adapt well to all clients and runtime execution contexts. The goal of our research is to develop methods and software to support *reconfigurable distributed applications* which can be customized to specific requirements. Thinking in terms of *software product line engineering* [2], we need a product line for a given Web application, each instance of which is for a specific execution platform and context. To achieve such a product line, we have to satisfy two requirements: universal accessibility and context-dependent component distribution.

2 Virtualization

We address both problems by using *virtualization* on execution platforms to provide a uniform programming environment for objects. Specifically, we have developed a virtual programming environment which hides the incompatibilities in different platforms and models. The environment provides a high-level abstraction: programmers no longer deal with pages, HTTP requests, cookies, sessions, which are entities in specific models and platforms; instead, they focus on the building objects and their interaction, and specifying the application logic. A Web application in the virtual environment represents a product line characterized by its application logic, not a specific implementation which can be deployed on a specific platform.

3 Separation of Concerns

Following the principle of *separation of concerns*, our virtual environment is not only platform independent but also location agnostic: there is no notions about object location and object movement. Nonetheless, such specifications are required to build an executable Web application. We designed a specification system including two parts: a specification language allowing developers write object distribution schemes, and tools to help the framework implementation

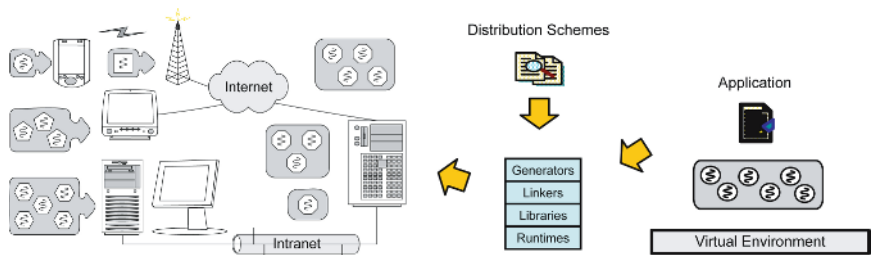


Fig. 1. System Architecture

execute the schemes. Combining an application with a specification, a Web application customized to a specific execution context can be defined. Inspired by the success of Web Style Sheets, our specification system is also based on *transformation*: converting a composition of objects into another composition of annotated objects. The annotation attributes are given by a specific execution framework. For example, in a typical Web framework, the location of an object may be a *client* or a *server*. Specifying object allocation can be achieved by annotating each object's with the desired attribute value. Sophisticated loading policies can be specified in a similar way. We have developed an algorithm to perform the transformation and proved its correctness.

4 Generative Approach

One approach to support virtualization over heterogeneous platforms is to define and build a middleware. This approach has the disadvantage of requiring the deployment of new software systems and protocols, which is impractical for the sheer scale of the Internet and the diversity of client platforms. Instead, we adopt the paradigm of generative programming [1] to realize application execution: before deployment, the composing objects of a Web application undergo a generative process to obey the annotations generated by the specification system.

To support these generated objects, a light-weight execution framework is required, whose primary role is to facilitate object management across HTTP channels. We implemented a Web execution framework composed of libraries and runtime systems. Our experience suggests that it is feasible to facilitate reconfigurable object distribution using a specification which controls object allocation and regulates loading patterns.

References

1. Krzysztof Czarnecki and Ulrich Eisenecker. *Generative Programming: Methods, Tools, and Applications*. Addison-Wesley, 2000.
2. Klaus Pohl, Gnter Beckle, and Frank van der Linden. *Software Product Line Engineering: Foundations, Principles, and Techniques*. Springer, 2005.

Towards the Definition and Validation of Coupling Metrics for Predicting Maintainability in Service-Oriented Designs

Mikhail Pereplechikov, Caspar Ryan, and Keith Frampton

RMIT University

School of Computer Science and Informational Technology
{mikhailp, caspar, keithf}@cs.rmit.edu.au

1 Coupling Metrics for Service-Oriented Designs

Service-Oriented Computing (SOC) is emerging as a promising software development paradigm based on the concept of encapsulating application logic within independent, loosely coupled, stateless services that interact via messages using standard communication protocols. The design principles of SOC are yet to be well understood; therefore service-oriented systems are often developed in an ad-hoc fashion, with little support for validating the quality of the service-oriented (SO) design upon which such systems are based. As such, there is a need for mechanisms supporting the measurement of internal structural quality attributes of SO designs (such as coupling and cohesion) in order to predict and enhance the quality of the final product.

Although there is a large body of research and practice on metrics for software developed using traditional approaches such as Object-Oriented (OO) and Procedural development, such metrics are not immediately applicable to SO systems. This is due in large to the following three reasons: i) SOC is based on *independent, platform agnostic services* that exhibit *high reuse* due to *low coupling* (coupling through service interfaces only); ii) *Services can be implemented by elements belonging to various development paradigms / languages*; iii) *SOC introduces an additional level of abstraction and encapsulation - a service*. In SOC, *operations* are aggregated into *elements* (classes/procedures, etc.) that implement the functionality of a *service* as exposed through its *service interface*.

As such, our current research is concerned with metrics for quantifying the structural *coupling* of design artefacts in service-oriented systems, where coupling is defined as a *measure of the extent to which interdependencies exist between software modules*. To date, a suite of design level metrics, covering all of the structural coupling aspects of a formal model of service-oriented design [4], has been derived and validated against the *property-based software engineering measurement* framework of Briand et al. [1]. These metrics are intended as early indicators/predictors of the quality characteristic of *maintainability* in service oriented systems, with the aim of enabling organisations to identify problems earlier in the software development life-cycle. The metrics are based on axioms which the authors' defined to establish connections between the *coupling* and *maintainability* of SO designs in terms of analysability, changeability, stability, and testability. The axioms are based on existing literature (e.g. [2]) and the authors' experience with service-oriented development [3].

There are ten axioms in total but due to space constraints only a single axiom and the definition of a single related metric are shown:

Axiom 1: High incoming coupling from service implementation elements¹ $e_1...e_n$ belonging to different services, to a given implementation element e of service s , will negatively influence *changeability*. This is because $e_1...e_n$ will be dependent upon the implementation characteristics of service s and thus the reuse of the external services containing these elements will be limited.

In total, nineteen metrics were derived and classified into three types: i) those directly supporting the axioms; ii) those included for coverage; and iii) those representing aggregations of the former two types of metrics. Below is an example metric based on the previously defined axiom 1. Note that the formal definition, motivation and validation of this metric have been omitted due to space limitations.

Metric Name: Weighted Extra-Service Incoming Coupling of Element (WESICE)

WESICE for a given service implementation element e of a particular service s is the *weighted* count of the number of system elements not belonging to the same service that couple to this element. To support the measurement process, a series of weights for different types of relationships defined in [4] was proposed based on the coupling effects described by the axioms. The weights are assigned in *two consecutive steps*. Firstly, the weights are assigned based on the types of elements involved in the communication. Secondly, the weights are assigned based on the locality and type of communication itself.

2 Future Work

The next step of this research involves empirical evaluation in order to test the axioms and quantify the effect of the metrics as predictors of *maintainability*. This testing will also facilitate the refinement of the metric weights. Following a successful outcome, ongoing work will involve the definition of metrics to measure complexity and cohesion. Finally, in the long term view, tool support for the automated collection of metrics will also be developed.

Acknowledgement. This project is funded by the ARC (Australian Research Council), under Linkage scheme no. LP0455234.

References

1. Briand, L.C., Morasca, S. and Basili, V.R. Property-Based Software Engineering Measurement. *IEEE Transactions on Software Engineering*, 22 (1), 1996. 68-86.
2. Erl, T. *Service-Oriented Architecture: Concepts, Technology, and Design*. Prentice Hall PTR, Indiana, USA, 2005.
3. Pereplechikov, M., Ryan, C. and Frampton, K., Comparing the Impact of Service-Oriented and Object-Oriented Paradigms on the Structural Properties of Software. in *Second International Workshop on Modeling Inter-Organizational Systems (MIOS'05)*, (Ayia Napa, Cyprus, 2005).
4. Pereplechikov, M., Ryan, C. and Frampton, K., A formal model of service-oriented design. in *13th Asia Pacific Software Engineering Conference*, (Bangalore, India, 2006), submitted for publication.

¹ OO classes/interfaces, procedural packages/headers, and business process scripts.

R&D Project Management with ODESeW

Asunción Gómez-Pérez¹, Angel López-Cima¹, M. Carmen Suárez-Figueroa¹,
and Oscar Corcho²

¹ OEG - Facultad de Informática. Universidad Politécnica de Madrid (UPM) Campus de
Montegancedo, s/n. 28660 Boadilla del Monte. Madrid. Spain
{asun, alopez, mcsuarez}@fi.upm.es

² University of Manchester. School of Computer Science. Manchester, United Kingdom
Oscar.Corcho@manchester.ac.uk

Abstract. ODESeW allows developing ontology-based Web portals. We describe the functionalities offered by a specific deployment of the ODESeW application development platform, oriented to the management of EU R&D projects. As an example of this specific deployment, we focus on the project management functionalities currently provided for the EU KnowledgeWeb¹ Network of Excellence (NoE).

1 Introduction

One important aspect in EU R&D project management is the periodic generation of reports about its state. Project reporting requires inputs from every partner in the consortium. Project coordinators are normally responsible for the generation and submission of the consolidated information, which has to be globally consistent.

We show how we ameliorate this problem by using the Semantic Web application development framework ODESeW (Semantic Web Portal based on WebODE) [1] to create a project portal that provides, among other functionalities, a set of project management functions. These functions are based on knowledge about project management and reporting that has been formalised by a set of ontologies.

ODESeW offers developers a set of services and tools for developing Semantic Web applications and gives, by default, navigation and visualization models that allow visualizing, editing and navigating the content in a portal. Such models can be modified and extended easily, permitting developers to create specific visualization and navigation models. The technical details are provided in [1].

2 Periodic Progress Reports in ODESeW

Here we focus on one specific deployment of ODESeW that is valid for setting up and maintaining the Web site of an R&D project, including the creation of periodic progress reports for EU R&D projects.

¹ Work supported by EU IST NoE Knowledge Web. <http://knowledgeweb.semanticweb.org/>

ODESeW includes a set of functionalities that help all members of the consortium, and particularly the coordinator, to generate and monitor management documents. On the one hand, it makes each partner be aware of the amount of information to be reported. On the other hand, it generates the activity and effort reports.

ODESeW can manage different domain ontologies. To describe R&D projects, we have used five project description ontologies (documentation, event, organization, person, and project), a user-role ontology (for managing different user profiles within the project), and a project management ontology (for managing R&D projects).

In the specific case study of Knowledge Web, the functionalities provided by the portal are divided according to the different users that can access it:

Partner user. This user is responsible of inserting his/her organization information and the information of all the participants in the project from his/her organization. If the partner is a WP leader, he/she is also responsible to upload deliverables inside the portal. Besides, this user can insert concrete meetings, conferences, workshops, etc.

Reporting users. the provided functionalities are oriented to guide the user through the different reports and the generation and submission of reports. The main tasks of this kind of user are:

- Workpackage (WP) progress reports, for each WP that the user's organization is leader of. This user must introduce the description of the progress done in a WP. Besides, the portal shows the user all the available deliverables that must be delivered in the current period report and all the delayed deliverables, so that the user can indicate their current status and the expected date of delivery.
- Effort report for the organization to which the user belongs. The technical annex of a project specifies the amount of effort that must be spent by each partner in different WPs during the overall project. In the management report the effort spent by partners in the different WPs during the period report should be shown.

Area managers. in this case, the functionalities are oriented to generate the area progress reports. In the context of Knowledge Web, WPs are organized in four areas: industrial, research, educational and management. In the activity report, area managers can include an overview about the general progress of the area.

Managing director. The managing director is a person that belongs to the project coordinator organization. The provided functionalities are oriented to monitor the evolution of all reports, to generate the complete reports and generate others management report concerning only to the coordinator partner.

Administrator. This user is in charge of creating new users, setting their read and write permissions and specifying which ontologies in the ontology server (WebODE) are being managed inside the portal. He/she is also allowed to change the ontologies inside the ontology server. Besides all administration issues, this user is in charge of including all the project definition information: WPs, deliverables, global efforts, etc.

Reference

1. López-Cima A, Corcho O, Gómez-Pérez A. 2006. A platform for the development of Semantic Web portals. In: Proceedings of the 6th International Conference on Web Engineering (ICWE2006). Stanford, July 2006.

Heavyweight Ontology Engineering

Frédéric Fürst¹ and Francky Trichet²

¹ LARIA - Laboratoire de Recherche en Informatique d'Amiens
CNRS-FRE 2733 - University of Amiens
UPJV, 33 rue Saint Leu
80039 Amiens Cedex 01 - France
frederic.furst@u-picardie.fr

² LINA - Laboratoire d'Informatique de Nantes Atlantique
CNRS-FRE 2729 - University of Nantes
2 rue de la Houssinière - BP 92208
44322 Nantes Cedex 03 - France
francky.trichet@univ-nantes.fr

Abstract. An *heavyweight ontology* is a *lightweight ontology* (*i.e.* an ontology simply based on a hierarchy of concepts and a hierarchy of relations) enriched with axioms used to fix the semantic interpretation of concepts and relations. Such an ontology can be a domain ontology, an ontology of representation, an ontology of PSM, etc. In our work, we argue in favor of using a graph-based solution to deal with the different activities related to Heavyweight Ontology Engineering, in particular ontology representation, ontology operationalisation, ontology evaluation (*i.e.* verification and validation) and ontology matching. Our approach consists in using the graph-based paradigm to represent all the components of an heavyweight ontology (*i.e.* Concepts, Relations and Axioms) and using graph homomorphism techniques to compare (at the conceptual level) the core components of an heavyweight ontology: the Axioms. This explicit graph-based representation of axioms coupled with reasoning capabilities based on graphs homomorphism facilitates both (1) the definition of important notions for Heavyweight Ontology Engineering such as *Compatible/Incompatible Axioms* or *Specialisation/Generalisation of Axioms* and (2) the topological comparison of axioms, which in our work is used to define a new approach of ontology matching mainly based on axiom-based ontology morphisms.

Keywords: Heavyweight Ontology, Axioms, Graph-Based Techniques, Ontology Matching, Ontology Evaluation, Conceptual Graphs.

1 Representing Heavyweight Ontologies with Conceptual Graphs

Axioms are the main building blocks for fixing the semantic interpretation of the concepts and the relations of an ontology, and this is what differentiates *lightweight ontologies* from *heavyweight ontologies*. Currently, there are not so many real-world ontologies that make substantial use of axioms. However, as introduced by T. Berners-Lee - "*For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning*" - we think that the need to develop *heavyweight ontologies* will

inevitably increase in an immediate future; this is also clearly demonstrated by the current W3C standardization efforts for a Semantic Web Rule Language.

To represent heavyweight ontologies, we propose OCGL (*Ontology Conceptual Graphs Language*) [2]. This modelling language is based on a graphical syntax inspired from those of the Conceptual Graphs model (CGs). The Conceptual Graphs model, first introduced by J. Sowa, is an operational knowledge representation model which belongs to the field of semantic networks. Two approaches for reasoning with CGs can be distinguished: (1) considered CGs as a graphical interface for logics and reasoning with logic and (2) considered CGs as a graph-based knowledge representation and reasoning formalism with its own reasoning capabilities. In our work, we adopt the second approach by using the projection (a graph-theoretic operation corresponding to homomorphism) as the main reasoning operator; projection is sound and complete w.r.t. deduction in FOL. OCGL has been implemented in a tool called TooCom (*a Tool to Operationalize an Ontology with the Conceptual Graph Model*) [1], which is available under GNU GPL license at <http://sourceforge.net/projects/toocom/>.

2 Reasoning on Heavyweight Ontologies with Conceptual Graphs

The explicit graph-based representation of axioms provided by OGCL coupled with reasoning capabilities based on graphs homomorphism facilitates the topological comparison of axioms at the ontological level. This feature (ontology morphism founded on graph-based knowledge representation and graph-based reasoning mechanisms) is interesting for several activities related to Heavyweight Ontology Engineering such as ontology matching or ontology evaluation.

In this context, we have defined (and implemented in TooCom) a new method for ontology matching [3]. This method mainly relies on the use of the axiomatic level of the ontologies to discover semantic analogies between concepts and relations. But our graph-based approach is not limited to ontology matching. Indeed, to generalize our principle of reasoning on heavyweight ontology with graph-based solutions, we have formally defined (in the context of the CGs) the notions of *Incompatible Axioms*, *Compatible Axioms* and *Specialisation/Generalisation of Axioms*. These notions can be used for multiple activities related to Heavyweight Ontology Engineering such as verification, validation, merging and evolution. We are currently using these notions to implement new functionalities of TooCom.

References

1. F. Fürst. TooCoM: a Tool to Operationalize an Ontology with the Conceptual Graph Model. In *Proceedings of the Workshop on Evaluation of Ontology-Based Tools (EON'2003) at ISWC'2003*, pages 57–70, 2003.
2. F. Fürst, M. Leclère, and F. Trichet. Operationalizing domain ontologies: a method and a tool. In R. Lopez de Mantaras and L. Saitta, editors, *European Conference on Artificial Intelligence (ECAI'2004)*, pages 318–322. IOS Press, 2004.
3. F. Fürst and F. Trichet. Axiom-based ontology matching. In *Proceedings of 3rd International Conference on Knowledge Capture (K-CAP'2005, Banff-Canada)*, pages 195–196. ACM Press. ISBN 1-59593-163-5, 2005.

Semantic Data Integration in a Newspaper Content Management System

A. Abelló¹, R. García², R. Gil², M. Oliva², and F. Perdrix^{2,3}

¹ Univ. Politècnica Catalunya
aabello@lsi.upc.edu

² Universitat de Lleida

{oliva, rgil}@diei.udl.es, roberto@griho.net

³ Diari Segre S.L.U.
perdrix@diarisegre.com

Abstract. A newspaper content management system has to deal with a very heterogeneous information space as the experience in the Diari Segre newspaper has shown us. The greatest problem is to harmonise the different ways the involved users (journalist, archivists...) structure the newspaper information space, i.e. news, topics, headlines, etc. Our approach is based on ontology and differentiated universes of discourse (UoD). Users interact with the system and, from this interaction, integration rules are derived. These rules are based on Description Logic ontological relations for subsumption and equivalence. They relate the different UoD and produce a shared conceptualisation of the newspaper information domain.

1 Introduction

From our experience in the newspaper content management systems domain¹, it has been possible to develop an experience of Semantic Web technologies in a real setting in the Diari Segre [1]. The main contributions are:

- An ontological framework for the newspaper domain [2].
- A semantic search and exploration user interface [3].

However, despite these achievements and as result of the experience acquired thereof, some aspects of this practical approach to a semantic newspapers have to be improved. Fundamentally, the main issue is the gaps among the different users' conceptual models and the ontologies that try to formalise a shared conceptual model.

Users of a newspaper system need to collect, to organise and to share lots of information about news. It is very difficult that different users classify a piece of news with the same topic, subject or keywords. This situation leads the users of the newspaper system to easily miss the needed information. The main problem with the Diari Segre newspapers content management system is the gap between journalists'

¹ This work is funded by the Spanish Ministry of Science and Technology, grants FIT-150500-2003-511, TIC2002-1948, and TIN2005-05406.

keywords and the topics used by archivist for classification. This gap prevents journalist from finding the content they need during their daily work. Thus, it makes them asking archivists to locate content for them, which is an overhead for them.

Altogether, this is a data integration problem caused by the interaction of different conceptual models. In order to overcome it, our approach is to formalise these conceptual models using ontologies. Once formalised, it is possible to employ computerised methods based on Description Logics in order to build up an integration service based on users' interaction.

2 Semantic Methodology for Data Integration

[4] argues for the importance of interactive and iterative integration. They present a tool that helps such process by looking at the instances to guide the conflict resolution at the schema level. In our case, we do not aim at solving any conflict, by assuming that they exist due to the different points of view of users. Thus, this will be handled by making their different points of view explicit, i.e. those known instances, also known as Universe of Discourse (UoD).

Regarding the criteria proposed in [5] to classify semantic integration approaches, ours would take the following values:

- Who generates the mappings: Agents themselves
- When define Agent-to-Agent mapping: Auto-generated at agent interaction time.
- Topology: Mediated.
- Degree of Agreement: Agree on subsumption/overlapping of Universes of Discourse (UoD).

Therefore, we have introduced a new value for the degree of agreement, and our proposal does not fit in any of the five architectures proposed there. Those values given to this attribute in [5] are all based on the ontologies, while we do not assume any agreement on the ontologies being used, but on the instances known by each user.

References

1. Diari Segre media group, <http://www.diarisegre.com>
2. García, R.; Perdrix, F. and Gil, R.: "Ontological Infrastructure for a Semantic Newspaper". In "Semantic Web Annotations for Multimedia Workshop, SWAMM 2006". 15th World Wide Web Conference, Edinburgh, UK, 2006
3. Castells, P.; Perdrix, F.; Pulido, E.; Rico, M.; Benjamins, R.; Contreras, J. and Lorés, J.: "Neptuno: Semantic Web Technologies for a Digital Newspaper Archive". Springer, LNCS Vol. 3053, pp. 445-458, 2004
4. Sattler, K.-U.; Conrad, S. and Saake, G.: "Interactive example-driven integration and reconciliation for accessing database federations". Information Systems 28(5), pp. 394-414, 2003
5. Uschold, M. and Grüninger, M.: "Architectures for Semantic Integration". In Kalfoglou, Y. et al (eds.): "Semantic Interoperability and Integration", Dagstuhl Seminar Proceedings, Num. 04391, Schloss Dagstuhl, Germany, 2005

Change Detection in Ontologies Using DAG Comparison

Johann Eder¹ and Karl Wiggisser²

¹ University of Vienna, Dep. of Knowledge and Business Engineering
johann.eder@univie.ac.at

² Klagenfurt University, Dep. of Informatics-Systems
wiggisser@isys.uni-klu.ac.at

1 Introduction and Motivation

An ontology is *an explicit specification of a conceptualization*[1]. Ontologies are seen as important technique for semantic data processing, and in particular for interoperability. As they represent knowledge about a certain evolving real world domain the ontologies have to evolve as well. Knowledge about the changes is mandatory to correctly interpret data or documents based on the semantics defined in the ontology. Furthermore, the correct comparison of data and documents from different points in time, based on different versions of an ontology is only possible if the differences between these versions are known.

Often, only the different versions of a changing ontology are available, but the change history is missing. To help in this situation is the ambition of the work presented here. In particular, we focus on the following problem: Given two versions of an ontology we want to derive a series of change operations, which is able to transform one version into the other and thus is an explicit representation of changes between the two ontology versions. As ontologies can be represented by graphs, we achieve this goal by means of graph comparison.

In [2] and [3] we presented a graph based approach for ontology versioning. Incorporating changes in such a temporal ontology is easy if the changes are known, but can be a very complex task, if the differences are unavailable. For instance, ontology development often is a federated task where individual teams work on different parts of the ontology. Integrating these parts without change information can be very hard. In other situations, the differences may be known, but tagging them for the versioning system is long lasting and error prone.

There are some other approaches for comparing ontologies. Among them Ontoview [4] and PromptDiff [5] are the best known ones.

2 Comparing Ontology Graphs

An ontology can be seen as a cyclic graph where the concepts are represented by nodes and the relations between concepts are represented by typed edges. For our approach we assume to have a *rooted DAG* (RDAG), i. e. a directed acyclic graph with only one node not having any parents, thus we have to transform

the ontology into such an RDAG. Such a transformation is always possible by introducing so called *slots*, which represent edges that may create cycles in the graph. On this RDAG, we define operations for inserting, removing and updating nodes, edges and slots.

For comparing two ontology versions, each version is represented by one RDAG. Then we apply our graph comparison algorithm, which delivers an edit script consisting of a series of operations, which transform one graph into the other. These operations represent the differences between the graphs and thus, the differences between the two ontology versions.

Our graph comparison is based on the tree comparison of Chawathe et al. [6] which we extended with renaming detection and the capability of comparing RDAGs [7]. We assume that ontologies do not change very much between two versions. The matching of concepts between versions is based on the fact that a concept's name often acts as key, thus duplicate names will seldomly occur and names will not change very often. If there is a renaming, it will be detected in a heuristic approach. After matching and renaming detection, the remaining differences can be found with two additional graph traversals.

We did an evaluation of our approach on randomly generated data, which showed surprisingly good results, both in terms of time and accuracy. Currently we are working on an evaluation against real world ontologies.

3 Conclusions

Ontologies represent knowledge from real world domains. As the real world tends to change, the ontologies also have to evolve. For incorporating such ontology changes into a versioning system, the differences have to be known. In this poster we outlined an approach for comparing two versions of an ontology by means of graph comparison which gave promising results after the first evaluation.

References

1. Gruber, T.: A Translation Approach to Portable Ontology Specification. *Knowledge Acquisition* 5(2) (1993) 1993
2. Eder, J., Koncilia, C.: Modelling Changes in Ontologies. In: Proc. of On The Move - Federated Conferences, OTM 2004, Springer (2004) LNCS 3292.
3. Eder, J., Koncilia, C.: Interoperability in Temporal Ontologies. In: Proc. of the Workshop on Enterprise Modelling and Ontologies for Interoperability. (2005)
4. Klein, M., Fensel, D., Kiryakov, A., Ogniyavov, D.: Ontology versioning and change detection on the Web. In: Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference. (2002)
5. Noy, N., Musen, M.: PromptDiff: A fixed-point algorithm for comparing ontology versions. In: Proc. of the Nat'l Conf. on Artificial Intelligence. (2002)
6. Chawathe, S., Rajaraman, A., Garcia-Molina, H., Widom, J.: Change detection in hierarchically structured information. In: Proc. of the ACM SIGMOD International Conference on Management of Data. (1996) 493–504
7. Eder, J., Wiggisser, K.: A DAG Comparison Algorithm and its Application to Temporal Data Warehousing. In: Proc. of the ER 2006 Workshops. (2006)

Filtering Internet News Feeds Using Ad-Hoc Ontologies

Lars Bröcker and Stefan Paal

Fraunhofer Institute for Media Communication
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
{Lars.Broecker, Stefan.Paal}@imk.fraunhofer.de

Abstract. News feeds using RDF Site Summary (RSS) are becoming common-place for news providers on the Internet. Using a news aggregator, users can stay informed about the topics they are interested in without having to constantly check the actual web pages themselves. This near instant availability of a plethora of news sources poses new challenges: filtering and categorisation techniques are needed that adapt to changing information needs of the users as well as to the changes occurring in the topics reported on. This paper presents an approach to filtering news feeds relevant to a set of interests using ad-hoc ontologies. These filters are created from a small core-ontology describing topics of articles the user is interested in.

1 Approach

Fig. 1 shows the workflow of our approach. As of yet, it is tuned to filtering news items in German only. This is due to characteristics of German, especially the abundance of compound-nouns. Step 1 collects the full texts of the articles from the RSS feed files.

Step 2 performs some pre-processing on the texts. They are tokenised in a per-sentence manner, stopwords are removed. This leaves only proper nouns and words containing upper-case letters, while preserving the positions of the words in the text. Finally, noun-chains are extracted, since these are deemed most interesting for ontology creation.

Step 3 collects the user-preferences. These are entered in a simple manner allowing the definition of classes, subclasses, properties, and instances. The goal is not to accurately model a part of the world but to seed the automatic process of ontology creation. The result of step 3 is a core-ontology formatted in an ontology language.

Step 4 takes two inputs: the list of noun-chains, and the core-ontology. First of all, the concepts of the ontology are checked for containment in the noun-chains. The chain-parts containing a concept are entered into the ontology as sub-classes of the concepts matched against whereas the whole chain is entered as an instance of this new subclass.

Step 5 performs a deeper analysis on the newly-created instances. The analysis focuses on the position of the subclass identifier in the word chain. This results

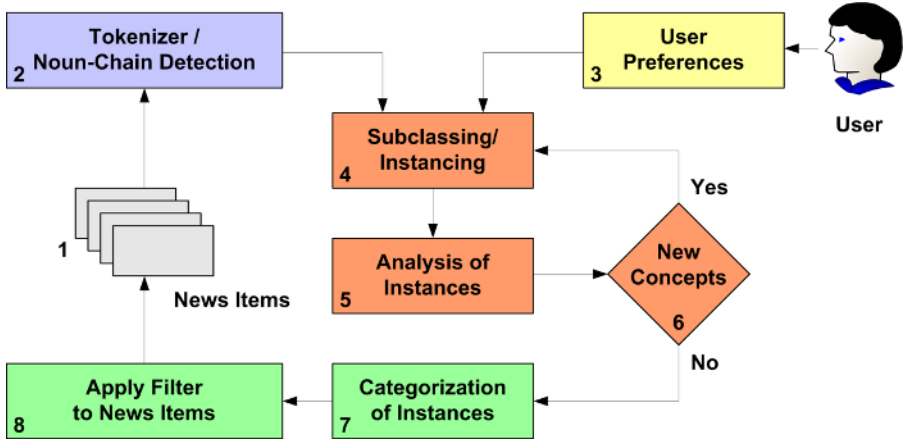


Fig. 1. Workflow of the Filtering Mechanism

in a list of candidates for inclusion into the ontology as new concepts. These are connected to the sub-classes they are derived from, so that their relationship is expressed. If over time a pattern emerges behind the automatically created relationships of concepts to a sub-class, the user can easily transform all of them at once.

Step 6 checks for new candidates and repeats steps 4 and 5 if that is the case.

Step 7 represents the post-processing stage. There, the properties entered by the user are evaluated and used to deduce categorisations aiding in transforming the string-instances into typed instances.

Step 8 performs the final operation in which the news items are annotated according to the ad-hoc ontology created by the prior steps. The news items are subsequently forwarded to the user, given that they contain at least one of the objects of the ontology.

2 Summary and Acknowledgments

We have conducted a couple of experiments using different core-ontologies and different news feeds. The results are promising. One experiment took a core-ontology comprised of 9 nodes as well as a document base of 400 articles. The resulting ontology had 428 nodes, 200 of which were instances. Of these 84.5% had been correctly categorised. This shows the feasibility of the approach with regard to the filtering of Internet news.

Our filtering system is able to quickly create topical filters, which reduce the amounts of incoming articles for the users and also perform some basic tagging on the articles they pass on.

This work is part of the project WIKINGER (see www.wikinger-escience.de), funded by the German Federal Ministry of Education and Research.

Mediation as Recommendation: An Approach to Design Mediators for Object Catalogs*

Daniela F. Brauner, Marco A. Casanova, and Ruy L. Milidiú

Department of Informatics – PUC-Rio
Rua Marquês de S. Vicente, 225 – Rio de Janeiro, RJ – Brazil
{dani, casanova, milidiu}@inf.puc-rio.br

Abstract. A catalog holds information about a set of objects, typically classified using terms taken from a given thesaurus. Mediated access to a collection of catalogs over the same domain therefore requires some strategy to deal with multiple thesauri, which represent different classifications for the same domain. This paper proposes an approach using online mapping rate estimations to define weighted relationships between terms of distinct thesauri. The mediator then uses such relationships to remap keyword-based queries to the different catalogs. Moreover, query answers provide valuable feedback to adjust the relationship weights, thereby improving the mediator accuracy.

1 Introduction

A *catalog* is a database that stores information about a set of objects, classified using terms taken from a given *object type thesaurus*. The design of a mediator for a collection of catalogs therefore requires aligning distinct object type thesauri, which is the central question we address in this paper.

We address this question by designing an online mapping rate estimator that gradually creates weighted relationships between terms of distinct thesauri by post-processing result sets returned by user queries. Briefly, given two catalogs, C_A and C_B , with thesauri T_A and T_B , if a query returns an object c from C_A classified as $t_a \in T_A$ and, again, c from C_B , but classified as $t_b \in T_B$, then c establishes some evidence that t_b maps into t_a . Note that this strategy depends on the assumption that the mediator can recognize when data from different catalogs represent the same object or not. For instance, in e-commerce applications the mediator may use the manufacturer's part numbers, if both catalogs store such information. Likewise, the mediator may use the object's spatial location in geographic information applications to try to deduce that two objects are indeed the same. The mediator then operates based on the online generation of evidences, and not on precisely defined thesauri alignments, which are very difficult to define a priori.

* This work is partially supported by CNPq under grants 550250/05-0, 140417/05-2 and 552068/02-0.

2 Estimating Relationships Between Terms of Two Thesauri

For simplicity, we assume that we have just two catalogs, C_A and C_B , storing objects from the same domain, classified using thesauri T_A and T_B , respectively. Also, we will be interested in mapping terms from T_A into T_B . However, the discussion can be generalized to bi-directional mappings and to more than two catalogs.

We say that entries $c_a \in C_A$ and $c_b \in C_B$ are *equivalent*, denoted $c_a \equiv c_b$, when they represent the same (real-world) object. The exact procedure that computes instance equivalence depends on the application.

A *user section* is a pair of queries Q_A over C_A and Q_B over C_B , submitted through the mediator. We assume that a user section contains queries that try to retrieve objects from C_A and C_B that are similarly classified. After the training process, the mediator will be able to recommend how to query C_B based on the terms used to query C_A .

Let $t_a \in T_A$ and $t_b \in T_B$ in what follows. The mediator maintains $P(t_a, t_b)$, the *mapping rate estimator* for t_a and t_b , which estimates the frequency that the term t_a maps to the term t_b . The mediator computes $P(t_a, t_b)$ as follows.

The mediator stores $n(t_a, t_b)$, the sum of the all occurrences of pairs of objects $c_a \in C_A$ and $c_b \in C_B$ such that: (1) $c_a \equiv c_b$; (2) the types of c_a and c_b are t_a and t_b , respectively; (3) c_a and c_b were observed in a previous user section. The mediator also stores $n(t_a)$, the sum of the all occurrences of objects $c_a \in C_A$ such that: (1) the type of c_a is t_a ; (2) c_a was observed in a previous user section.

The mediator post-processes the result sets a new user section and computes $\Delta n(t_a, t_b)$ and $\Delta n(t_a)$, defined exactly as $n(t_a, t_b)$ and $n(t_a)$, except that the objects are those in the result sets of the new user section. Then, the mediator recomputes $P(t_a, t_b)$ as follows:

$$P(t_a, t_b) = \frac{\Delta n(t_a, t_b) + \alpha \cdot (n(t_a, t_b) + \Psi)}{\Delta n(t_a) + \alpha \cdot (n(t_a) + I)}$$

where

α is a coefficient that takes values from the set $\{0.01, 0.1, 0, 1, 10, 100\}$, calibrated during the model validation process

$\Psi = \frac{I}{|T_B|}$ is a smoothing coefficient assumed as the inverse of the size of the thesaurus of the second term.

Note that the above equation is symmetric in t_a and t_b and can be easily adapted to compute estimations for the frequency that the terms in T_B map into terms in T_A .

To validate and test the estimation model, we used the ADL Gazetteer (<http://www.alexandria.ucsb.edu/gazetteer>) and the GEOnet Names Server (<http://gnswww.nga.mil/geonames/GNS>). We stored data from both gazetteers locally and partitioned the data into a tune set and a test set. We applied the 6-fold cross-validation technique to calibrate the model parameters. As a result, we obtained 26 pairs of terms from T_A to T_B aligned with mapping rate greater than 0.4, with accuracy of 89.7% and recall 81.3%. From T_B to T_A , we obtained 44 pairs of terms aligned with accuracy of 93.6% and recall of 95.7%.

Benchmarking Data Schemes of Ontology Based Databases

Hondjack Dehainsala, Guy Pierra, and Ladjel Bellatreche

Laboratory of Applied Computer Science (LISI) ENSMA - Poitiers University, France

Nowadays, ontologies are more and more used in several research and application domains, in particular e-commerce, e-business and the Semantic Web. Several tools for managing ontology data and ontology-based data (also called individuals or ontology class instances) are available. Usually, ontology-based data manipulated by these tools are stored in the main memory. Thus, for applications that manipulate a large amount of ontology-based data, it is difficult to ensure an acceptable performance. So, there is a real need for storing ontology-based data in database schemas to get benefit of the capabilities offered by DBMSs (query performance, data storage, transaction management, etc.). We call this kind of databases Ontology-Based Databases (OBDBs).

Over the last five years, two main OBDB structures for storing ontology and ontology-based data were proposed [1]. In the single table approach, the description of classes, properties and their instances are described by means of *triples* (subject, predicate, object) stored in a single table, called the *vertical table*. In the dual scheme approach, ontologies are described by specific schemas, depending upon the ontology model, but instances are stored either as set of *triples* in a single vertical table, or in a set of unary and binary tables, with one table per class, representing the identifiers of its instances, and one table per property, representing the pairs (instance identifiers, property value), this structure is known as the *decomposition model*. Unfortunately, all these approaches are poorly adapted when ontology-based data contains a large number of instances described by many properties values. In this case, any query requires a large number of join operations. This kind of ontology-based data is largely used in several application domains, particularly in e-commerce and e-engineering (which are our two main application domains). We have proposed a new architecture of OBDB, called OntoDB (Ontology Data Base). In this paper, we present the main characteristics of our approach, and the results of the benchmark used to compare our new structure with the existing structures. This benchmark uses a real ISO-standardized ontology for electronic commerce.

1 Overview of the OntoDB Model

Our approach requires that ontology-based data fulfil three requirements: (1) each property is defined in the context of a class that specifies its *domain* of application, and it is associated with a *range*, (2) all the classes to which an instance belong have exactly one minimal class for the subsumption relationship, this class is the instance *base class* and (3) only properties that are *applicable* for

its base class may be used for describing an instance. Note that these requirements are fulfilled in a number of cases including e-commerce data. Then the OntoDB model consists of four parts [2]. The *ontology* and *meta-schema* parts are used for storing respectively ontologies data, and ontology models within a reflexive meta-model. The *meta-data* part is the usual database catalogue that describes the table structure of the three other parts. Finally, the *data* part is used for storing ontology-based data. Unlike in classical approaches, in OntoDB ontology class instances are stored using a *table per class* approach. It consists in associating a table to each ontology class. Columns of this table represent those rigid properties of the class that are associated with a value for at least one of its instance. Property values of each instance are represented in the same row.

2 Benchmarking OBDB Models

Our benchmark is based on a real ontology standardized as IEC 61360-4. This ontology describes the various kinds of electronic components together with their characteristic properties. It is composed of 190 classes with a total of 1026 properties. We have generated automatically various sets of ontology-based data by varying the number of instances per classes and the number of properties values per instance. The size of test data falls within the range 0.3 GB-4 GB. Our test was based in three kinds of queries: (1) targeted class queries, where the user is supposed to know the root class of the subsumption tree to be queried; (2) non targeted class queries, where the user does not know what kind of ontology class she is looking for, and (3) insertion and update queries. Details of all our tests and results are presented in [3]. For queries (1) and (3), our table per class approach outperforms the two classical approaches as soon as more than one property value is requested. As a rule, the ratio is bigger than 10. The only case where the decomposition model is better than our approach is for the non targeted class queries when the user requests a very small number of property values. We note that this kind of queries nearly never happens in our application domain. Engineers always knows what they are looking for before searching for property values.

Our conclusion is that the OntoDB approaches outperforms all the classical approaches for processing in particular e-commerce data.

References

1. V. C. A. Magkanaraki, S. Alexaki and D. Plexousakis. Benchmarking rdf schemas for the semantic web. In *Inter. Semantic Web Conference (ISWC)*, 2002.
2. L. Bellatreche, G. Pierra, X. Nguyen, H. Dehainsala, and Y. Ait Ameer. An a priori approach for automatic integration of heterogeneous and autonomous databases. *Inter.Conf. on Database and Expert Systems Applications*, (475-485), 2004.
3. H. Dehainsala, G. Pierra, and L. Bellatreche. Managing instance data in ontology-based databases. Technical report, LISI-ENSMA, <http://www.lisi.ensma.fr/ftp/pub/documents/reports/2006/2006-LISI-003-DEHAINSALA.pdf>, 2006.

Dynamic Load Balancing for Online Games Using Predefined Area Information

Beob Kyun Kim, Dong Un An, and Seung Jong Chung

Dept. of Computer Engineering, Chonbuk National University, South Korea
{kyun, duan, sjchung}@chonbuk.ac.kr

Abstract. Usually, online games cannot adapt to the dynamic change of population. In this paper, we propose a new dynamic load balancing approach for online games using predefined area information. To control dynamic change of population, it divides or merges fields considering the surveyed load based on predefined area information. By the simple modification to this information, we can easily control problems that come from changes of map data, changes of resources' status, or changes of user's behavior pattern.

1 Predefined Scenario Information Enabled Online Game

Several concepts [2] [3] [4] and partitioning algorithms [5] have been proposed to limit the number of neighboring avatars. But usually, they cannot survive from dynamic change of population and they don't consider the special relationships between neighbored areas. Each area has different properties, which are related to neighbored areas' properties and have different power, to decoy players.

We design a new dynamic load balancing approach using predefined area information which consists of Fields, Cell Groups, and Cells. Field is an area which is controlled by single game server. And a field consists of several cell groups. The system tries to divide or merge fields based on the load of its cell groups. A cell group is defined as a set of cells considering relationships between neighbored areas. Cell group is the smallest unit for partition and incorporation. Cell is the minimum unit to define cell group and different number of cells can be used to define cell group. Predefined area information has tree-like hierarchical structure. The coverage of a parent node A is same to that of union of its children. After partition of a field, new fields are subsets of this field.

2 Dynamic Load Balancing Using Predefined Area Information

The proposed system is follows traditional architectures based on networked servers. Balancing server is the only new component. It surveys loads of each field servers, checks if the surveyed load is affordable, and tries to divide or merge fields.

To control dynamic change of population, the balancing server monitors the surveyed load. If the surveyed load of a field exceeds the high limit, it chooses two subsets within all subsets of this field. Each subset is a set of cell groups. On the contrary,

if the load of a field not reached the low limit, the merge process works using dividing history. From the experimental results, our approach achieves about 23~67% lower loads for each field server. By the simple modification to this predefined area information, we can easily manage problems that come from changes of map data, changes of resources' status, or changes of user's behavior pattern.

Overheads, which are not appeared in the static management, may be produced by the monitoring process and the partitioning process. Usually in online games, the load of each field server is monitored in real-time and is used in other analysis process. The balancing server's intervention between monitoring and recording process is the only difference and its overhead is very trivial. The overhead produced by the partitioning process may include the synchronization of avatar, map data loading for newly divided or merged field, and rerouting of client and daemon server to new field. The synchronization of PC information is basic in online games and the only one more pass is added in the proposed system. And because the map data is static, each field server can have the entire map and must have different active area. Usually, the rerouting of client and daemon server to new field is the same process as the rerouting by avatar's migration. The only difference from the migration process is the number of PC and it can be improved by the improvement of the architecture or the migration algorithm.

3 Conclusions

In this paper, a dynamic load balancing approach for online games using predefined area information is proposed. To control dynamic change of population, it divides or merges fields based on predefined area information. From the experimental results, our approach achieves about 23~67% lower loads for each field server. By the simple modification to this predefined area information, we can easily manage dynamic change of population.

References

1. J.C.S. Lui, M.F. Chan.: An efficient partitioning algorithm for distributed virtual environment systems. *IEEE Transactions on Parallel and Distributed Systems*, Vol. 13, No. 3. (2002) 193-211
2. S. Singhal, M. Zyda.: *Networked Virtual Environments*, ACM Press, New York, (1999)
3. D.B. Anderson, J.W. Barrus, J.H. Howard.: *Building multiuser interactive multimedia environments at MERL. Multimedia*, Vol. 2, No. 2. *IEEE*. (1995) 77-82
4. C. Greenhalgh.: Awareness-based communication management in the MASSIVE systems. *Distrib. Syst. Engng* 5. UK. (198) 129-137

Seamless Integration of Generic Bulk Operations in Grid Applications

Stephan Hirmer¹, Hartmut Kaiser¹, Andre Merzky², Andrei Hutanu¹,
and Gabrielle Allen¹

¹ Center for Computation & Technology, Louisiana State University
² Vrije Universiteit, Amsterdam

Objectives

Within grid environments, as the number of remote operations increases, the latencies for these operations become a dominant factor for application performance. Bulk operations represent one potential methodology to reduce latencies for large numbers of similar operations. However, the identification of bulks can pose a non-trivial exercise for application developers, often implying changes in the remote API, and hence direct code modifications to the application.

In this work, we identified those properties an API must possess to provide for transparent bulk optimizations, allowing a latency hiding approach which does not require code changes on application level, and without complicating the remote API itself. We have developed a prototype implementation (within the SAGA¹ C++ reference implementation effort), and present performance measurements for bulks of remote file copy operations.

API Properties for Bulk Optimization

Three requirements which, when met, allow for bulk optimizations in remote API implementations without changing the API are: explicit asynchronicity, implicit concurrency information, and inspectability of asynchronous calls.

Explicit Asynchronicity: The main problem of transparent support for bulk optimization is to identify those asynchronous operations which can be collectively executed as bulks: *APIs must have explicit expressions for asynchronous operations.*

Implicit Concurrency Information: Only those asynchronous operations which have no dependencies from each other can be collected in bulks: *concurrency information of the asynchronous operations must be available, either explicitly or implicitly*².

Inspectability of Asynchronous Operations: To form a bulk operation, the operations to be clustered must be semantically similar (e.g., they must all be

¹ Simple API for Grid Applications.

² Asynchronous operations are considered concurrent if the applications behaviour is not altered when the execution order for these operations is changed.

read operations on the same file). Thus, *semantic information must be available* to the API implementation.

Properties of the SAGA API

The Simple API for Grid Applications, SAGA [1], is a standardization effort within the Open Grid Forum (OGF), which strives to enable applications to make use of grid programming paradigms. The support of bulk operations is an explicit design objective for the SAGA API.

In SAGA, a `saga::task` instance represents an asynchronous procedure. Multiple tasks can be collectively controlled in a task container. According to the above requirements, bulk optimizations can be transparently implemented within SAGA:

- tasks are an explicit expression of asynchronous operations on API level;
- tasks in a task container are, by definition, concurrent;
- tasks contain the complete semantics of the asynchronous operation.

The existence of task containers in the SAGA specification provides the means to identify independent tasks, and to optimize their execution.

Results

We implemented generic bulk optimizations for the SAGA C++ reference implementation, without introducing any changes to the SAGA API. The bulk optimization is applied to all operations, but is only effective if it is supported by the middleware – otherwise it falls back to normal one-by-one execution.

Our results confirm that bulk operations perform better than serial execution [3]. However, the implementation adds up to a 14% overhead for transparent bulk detection.

Figure 1 shows the elapsed wall time for a 3rd party file transfer of (1...500) files (1 MByte each) with the GridLab File Service [2], both for direct invocation of the service, and for indirect invocation via the SAGA call stack: task analysis, bulk creation, adaptor invocation, and service invocation. The used LAN was not isolated, and shows random traffic.

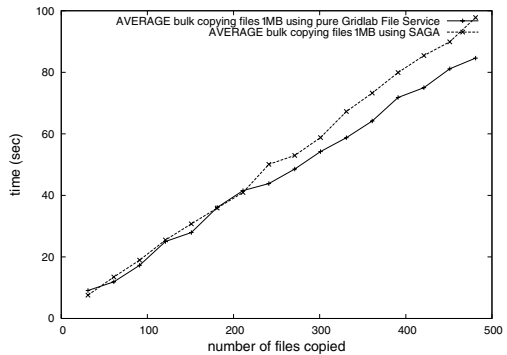


Fig. 1. Benchmark results: direct middleware vs. SAGA initiated bulk operations

References

1. T. Goodale, S. Jha, H. Kaiser, T. Kielmann, P. Kleijer, G. von Laszewski, C. Lee, A. Merzky, H. Rajic, and J. Shalf. SAGA: A Simple API for Grid Applications – High-Level Application Programming on the Grid. *Computational Methods in Science and Technology: special issue "Grid Applications: New Challenges for Computational Methods"*, 8(2), SC05, November 2005.
2. GridLab FileService. http://www.gridlab.org/WorkPackages/wp-8/file_service/.
3. E. Walker, R. Floyd, and P. Neves. Asynchronous remote operation execution in distributed systems. *Distributed Computing Systems, 1990. Proceedings., 10th International Conference on*, pages 253–259.

An Efficient Search Scheme Using Self-organizing Hierarchical Ring in Unstructured Peer-to-Peer Systems*

Saeyoung Han, Jaeui Sohn, and Sungyong Park

Dept. of Computer Science and Interdisciplinary Program of Integrated Biotechnology
Sogang University, Seoul, Korea
{syhan, andrew, parksy}@sogang.ac.kr

Abstract. We propose an efficient search scheme based on a self-organizing hierarchical ring structure in unstructured peer-to-peer systems. Our solution maintains a consistent super-peer ratio and makes the peers with relatively high capacities super-peers in dynamic environments. The benchmarking results show that the proposed algorithm outperforms the static algorithm that has the fixed number of rendezvous peers such as JXTA.

In hierarchical unstructured peer-to-peer networks utilizing peers' heterogeneity, super-peers with larger capacity form a super-peer network and serve their neighbors by storing the peers' advertisements and forwarding the queries from their neighbors through flooding, random walk, or routing using DHT. While using DHT in rendezvous network makes the search in JXTA faster, there is no control mechanism of the number of super-peers when the network size is changed or the peers' capacities are skewed in JXTA.

Our solution provides an efficient P2P search algorithm based on a self-organizing hierarchical ring that maintains a consistent super-peer ratio in dynamic environment and keeps the peers with relatively high capacities as super-peers. This consists of four algorithms, agent population and distribution control algorithm, peer selection algorithm, query routing algorithm, and the Gossip algorithm.

First, all peers collect the capacity information of neighboring peers via *mobile agents*. In order to reduce the excessive overhead by agents and collect sufficient information under dynamic peer-to-peer network, we should control the agent population appropriately. When an agent arrives at a peer, it computes the elapsed time, t , since the last agent arrived at the peer. If t is less than a termination threshold t_{th} , the agent terminates itself. But if t is greater than a cloning threshold t_{cl} , a new agent is generated with probability p_{cl} and migrated to another peer as well as the previous one. The probabilistic cloning decision prohibit the excessive agent generation. If t is between t_{th} and t_{cl} , the agent is migrated another peer based on the migration strategy. Under our migration strategy, each agent is likely to select the neighbor peers with the larger inter-arrival time of agents and smaller number of degree as a migration destination.

Second, based on the capacity information of other peers collected by agents, each peer estimates the average capacity of the whole system. The peer whose capacity is ρ

* This work was supported by the Brain Korea 21 project in 2006.

times greater than the average increases its own counting variable, where ρ is determined by the ratio of super-peers. Otherwise, the peer decreases its counting variable. Using two thresholds for promotion and demotion, each super-peer decides to demote or not, and each leaf-peer decides to promote or not at every iteration.

Third, unlike other hierarchical systems where peers directly reach super-peers, in order for peers to publish advertisements or send query requests through super-peers, each peer first have to look for any super-peer via random walk. When a super-peer is found, it directs the request to a destination super-peer using hash table. But if the super-peer that should store the requested advertisement does not store the advertisement, it forwards the request to its super-peer neighbors next to it on the super-peer view until the TTL is expired or find the super-peer that stores the advertisement. We call this process the super-peer walk.

Last but not least, the super-peers in our system utilize the *gossip algorithm* [1] as a membership protocol to maintain loosely-consistent super-peer view. When the number of super-peers attending the *gossip* is N , at least $\log(N) + c$ of super peers will be fanout. Therefore, each super-peer generates *gossip* messages for the $\log(N) + c$ peers among the peer view, and then sends them to $\log(N) + c$ of randomly chosen super-peers. On receiving *gossip* messages, the super-peers update their peer view with the peer view information inside the messages.

Using an event-driven simulator, we compared the performance of our self-organizing ring algorithm with that of JXTA when the size of the system is gradually increased. As we can see from Fig.1 (a), the number of success messages of our approach is larger than that of JXTA. This indicates that the peer view synchronizing protocol manages to handle the *churn* situation of super-peers in our approach. Moreover, Fig.1 (b) shows that the average search time of success messages in our approach is better than that in the JXTA. The search time depends largely on not only the traverse hop count but also the peers' capacity. The self-organizing super-peer ring algorithm allows the peers with higher capacity to be promoted as super-peers and causes the peers with lower capacity to be demoted as non-super-peers continuously. As a result of that, our approach can handle more messages faster than the JXTA system with fixed member of super peers. For the overhead, our approach has more overhead messages than JXTA system due to the additional messages related to agents and the membership protocols as shown in Fig. 1 (c).

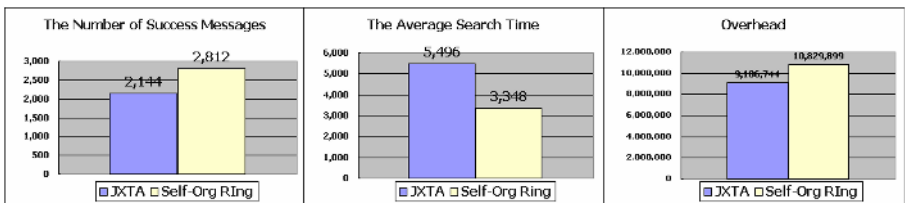


Fig. 1. (a) Number of Success Messages (b) Average Search Time (c) Overhead

Reference

1. Ganesh, A.J., Kermarrec A.M., Massoulie, L.: Peer-to-Peer Membership Management for Gossip-Based Protocols. IEEE Transactions on Computers (2003)

AWeSOMe 2006 PC Co-chairs' Message

The Second International Workshop on Agents, Web Services and Ontologies Merging (AWeSOMe 2006) is held in conjunction with the OnTheMove Federated Conferences (OTM 2006) in Montpellier, France, October 29, 2006. AWeSOMe is an interdisciplinary workshop focusing on research and applications combining Web services, ontologies and agents leading to the development of an intelligent service Web.

Web services are a rapidly expanding approach to building distributed software systems across networks such as the Internet. A Web service is an operation typically addressed via a URI, declaratively described using widely accepted standards, and accessed via platform-independent XML-based messages.

Emerging ontologies are being used to construct semantically rich service descriptions. Techniques for planning, composing, editing, reasoning about and analyzing these descriptions are being investigated and deployed to resolve semantic interoperability between services within scalable, open environments.

Agents and multi-agent systems can benefit from this combination, and can be used for Web service discovery, use and composition. In addition, Web services and multi-agent systems bear certain similarities, such as a component-like behavior, that can help to make their development much easier.

The Program Committee members did an excellent job in selecting the papers. All submitted papers underwent a thorough review process with each paper having at least two out of 26 submissions for the presentation at the workshop.

We would like to thank the members of the Program Committee who gave their time and energy to ensure the high quality of the technical program. We extend many thanks to our invited speaker, Barry Norton, for the generous sharing of his expertise. We are grateful to the OTM 2006 organizers for their support and encouragement. Especially, we would like to thank the OTM 2006 Chairs, Robert Meersman and Zahir Tari.

We acknowledge the efforts of the authors of selected papers for their contributions to the new and exciting interdisciplinary area covered by the workshop.

August 2006

Daniel Grosu, Wayne State University, USA
Pilar Herrero, Universidad Politécnica de Madrid, Spain
Gonzalo Médez, Universidad Complutense de Madrid, Spain
Marta Sabou, The Open University, UK

3-Level Service Composition and Cashew: A Model for Orchestration and Choreography in Semantic Web Services

Barry Norton and Carlos Pedrinaci

Knowledge Media Institute, Centre for Research in Computing,
Open University, Milton Keynes, UK
{b.j.norton, c.pedrinaci}@open.ac.uk

Abstract. There are two types of behavioural model in the WSMO semantic description of services: an orchestration and a choreography, together called the interface. While an orchestration defines a service's behaviour as a composition of existing parts, a choreography is intended to document the conversation of messages exchanged with a single client. In this paper we present a three-level model for behavioural descriptions, and how the Cashew workflow model fits into this, building on existing work in, and establishing connections with, semantic web services, workflow, and software engineering design.

1 Introduction

Cashew is an ontological model for workflow-oriented descriptions of semantic web service interfaces, descended from an earlier generalised representation of the OWL-S process model [10], called CASheW-S [22]. The definition of Cashew has extended, from this previous work, to accommodate WSMO [8] in three ways:

- whereas previously only orchestrations were described in workflow terms, with primitive choreographies being implicitly described via automata, now choreographies are also given a high-level description;
- whereas previously the 'unit of composition' was an operation on a service, now orchestrations compose goals, and choreographies compose operations;
- whereas previously dataflow connections were simple, specifying only performance outputs as sources and performance inputs as targets, now these are specified as types of WSMO mediators.

As well as this alliance with WSMO, the redesign of Cashew has attempted to build stronger links with two other communities. Firstly, the workflow control forms used have been re-examined in the context of 'workflow patterns', where the aim is to standardise the vocabulary for workflow in business process management [26]. Secondly, for the visual representation of workflow models we have looked at UML, which aims at a standard language for software design [17]. We consider all of these as background work in the following section, propose our language in Section 2, consider its representation in UML in Section 3 and then conclude and consider future work in Section 4.

1.1 OWL-S

The process model in OWL-S is an algebra of workflows, called processes, where the atomic processes are grounded to operations on web services. Although this has been called ‘service composition’, we have previously pointed out [20] that this is really a model of ‘operation composition’, *i.e.* a composite process is used to define a single operation, not a general service with multiple operations, and within that definition, the attachment to services of operations is not considered.

The gap between services and operations widens further when we consider statefulness of interaction in the service interface. When we allow that there is a ‘protocol’ governing the order of use of operations of a service in any given session, it becomes important that we consider the notions of service and session, and also that we give a semantic model to this protocol.

It is for these reasons that we consider that the OWL-S process model is only a useful model for orchestrations formed over stateless services, but we shall see that it is deficient when we must deal with services with protocols.

1.2 WSMO

The fragment of the WSMO meta-model [8] we deal with is represented, as a UML class diagram, in Figure 1. The central concepts, duly shown centrally, are ‘web service’, ‘mediator’ and ‘goal’.

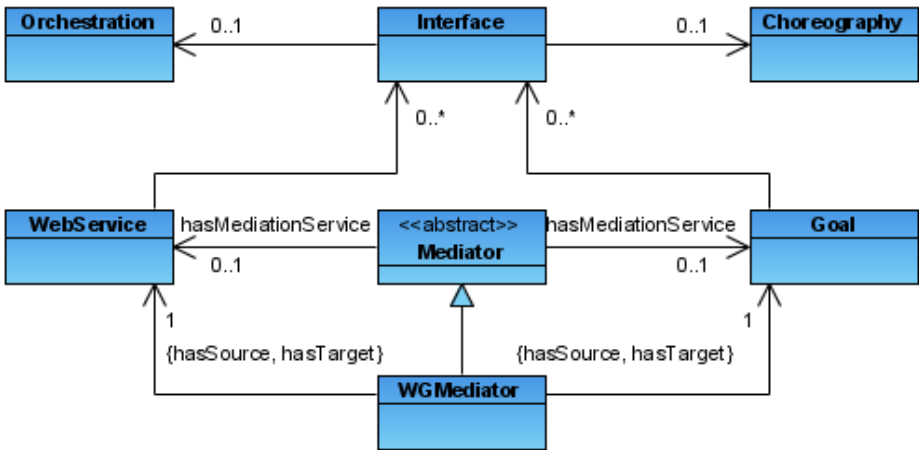


Fig. 1. (Partial) WSMO Meta-model

An instance of ‘web service’ is the basis for a semantic description of a web service, divided into two parts: the ‘capability’, not shown or considered in this paper, contains the *functional* description of the service; and the ‘interface’, which contains *behavioural* descriptions.

Following the connections from interface, we see that each instance here can contain an orchestration and/or a choreography. We also see that the description of a goal may contain interfaces, so that goals are more than simply first class representations of what have been called ‘service templates’, *i.e.* only functional descriptions of service requirements, following the initial publication of OWL-S.

So far in WSMO, the formal basis of orchestration and choreography in the form of abstract state machines (*ASMs*) [2] has been prescribed. Thereafter use of *ASM* choreographies to define the protocol that a client must use in order to consume a service has been defined [9] and exemplified [11]. The use of *ASMs* for orchestration is less developed [7].

Even the restricted two-party view on the definition of choreography that WSMO takes¹ is open to refinement. The DIP project² has proposed [15] that the distinction that the IRS [4], one of two DIP implementations of the WSMO model, makes between *service-* and *client-choreographies* [5] should be formalised in WSMO. The first difference between the two is the viewpoint: in a service choreography, the service documents which communications are offered to clients; in a client choreography, the client documents which communications they offer and accept in return. Although to some degree orthogonal, the second difference is that a client choreography will be finite and contain only those communications intended to achieve some goal, whereas the service choreography will document all allowed interactions with the service, which may be infinite.

The latter form, which DIP calls service choreography, is what is documented in the Amazon Case Study [11], and shows an infinite behavioural model where the client can interact with the Amazon service by searching at any point, and interleave this with logging in, logging out and making purchases. An example of a client choreography might connect a goal ‘buy the most recent edition of book x by author y ’ to this service via the refined conversation consisting of search, followed by log-in, purchase and log-out. Due to this goal-orientation of client choreographies, the meta-model in this paper will suggest, as detailed in Section 2, that client choreographies are actually choreographies attached to *wg*-mediators, a specialisation of mediators that, as shown in Figure 1, are used to attach goals to services.

The meta-model proposed in this paper will also follow IRS, in turn following the literature in problem-solving methods, in insisting that the choreographies of goals are always ‘one-shot’, *i.e.* that in expressing the desire to achieve a goal, a user does not have to worry about ‘control flow’, or complex interaction, and merely present some inputs and waits for an output. We will also follow the IRS in viewing orchestrations as a composition of goals, *i.e.* the units of composition in an orchestration will be abstracted to this level, and goal-based invocation will be used to match these goals to service at run-time, which may involve the use of discovery. In this way, the use of several operations of the same service can be abstracted into a goal that may then be considered atomic for composition.

¹ Which can be contrasted with the multi-party viewpoint of W3C [13], but also seen as a projection on this.

² <http://dip.semanticweb.org>

1.3 Workflow Patterns

‘Workflow Patterns’ are the result of a long-running project to formalise the possible variations between workflow systems, and provide a common vocabulary to compare these [26]. It is telling that, when presented³, one of the examples given of hidden differences between workflow systems in the early days was the distinction between, in workflow patterns terms, ‘XOR’ and ‘Deferred Choice’. In general terms, when asked whether their systems supported a choice operator, vendors would answer ‘yes’. On the other hand, given the vocabulary to ask whether systems supported choices resolved internally by the engine and choices resolved externally by the outcome of component tasks, the answer was too often only ‘yes’ one or the other, rather than both. In fact, this very example is another reason for the deficiency of OWL-S in the presence of choreography, since deferred choice is not supported by OWL-S. We shall show the extension to, and use of, this form of choice operator in the Cashew model.

1.4 UML

The UML is an ongoing effort by the Object Management Group⁴, standardised by ISO [17], to provide a language for software engineering design via diagrams describing both the static and dynamic characteristics of software artifacts. Of particular relevance are Activity Diagrams which, it has been suggested, are expressive enough to represent visually many of the Workflow Patterns [6] [27].

2 The Cashew Model

An extended ‘3-level’ WSMO meta-model, as proposed in the DIP project, is diagrammed in Figure 2. As sketched above, the concepts of both orchestration and choreography are abstracted, so that ASMs form just one way to represent these. The upper two levels represent a high-level view on these behavioural models, one oriented towards first-class workflow features and another towards diagramming for human inspection. In the DIP project these levels are filled by the Cashew workflow language and by UML Activity Diagrams. These are not the only possibilities, however, and in Section 4 we discuss others.

The lower half of the diagram documents a proposed extension to WSMO that has already been abstracted out from the DIP work and proposed to the working group [21]. The notion of ‘Performance’, due to OWL-S, allows us to hierarchically compose workflows from *performances* of workflows, so that each defined workflow can be reused in other contexts than where it was defined, and each instance is given a different identifier in context.

The details of the Activity Diagram meta-model — called ‘ADO’, the Activity Diagram Ontology — are not reproduced here; the reader is directed to the annex [12] of the relevant DIP deliverables.

³ Wil van der Aalst’s ‘Life After BPEL?’, presented as keynote at WS-FM’05.

⁴ <http://www.omg.org/technology/documents/formal/uml.htm>

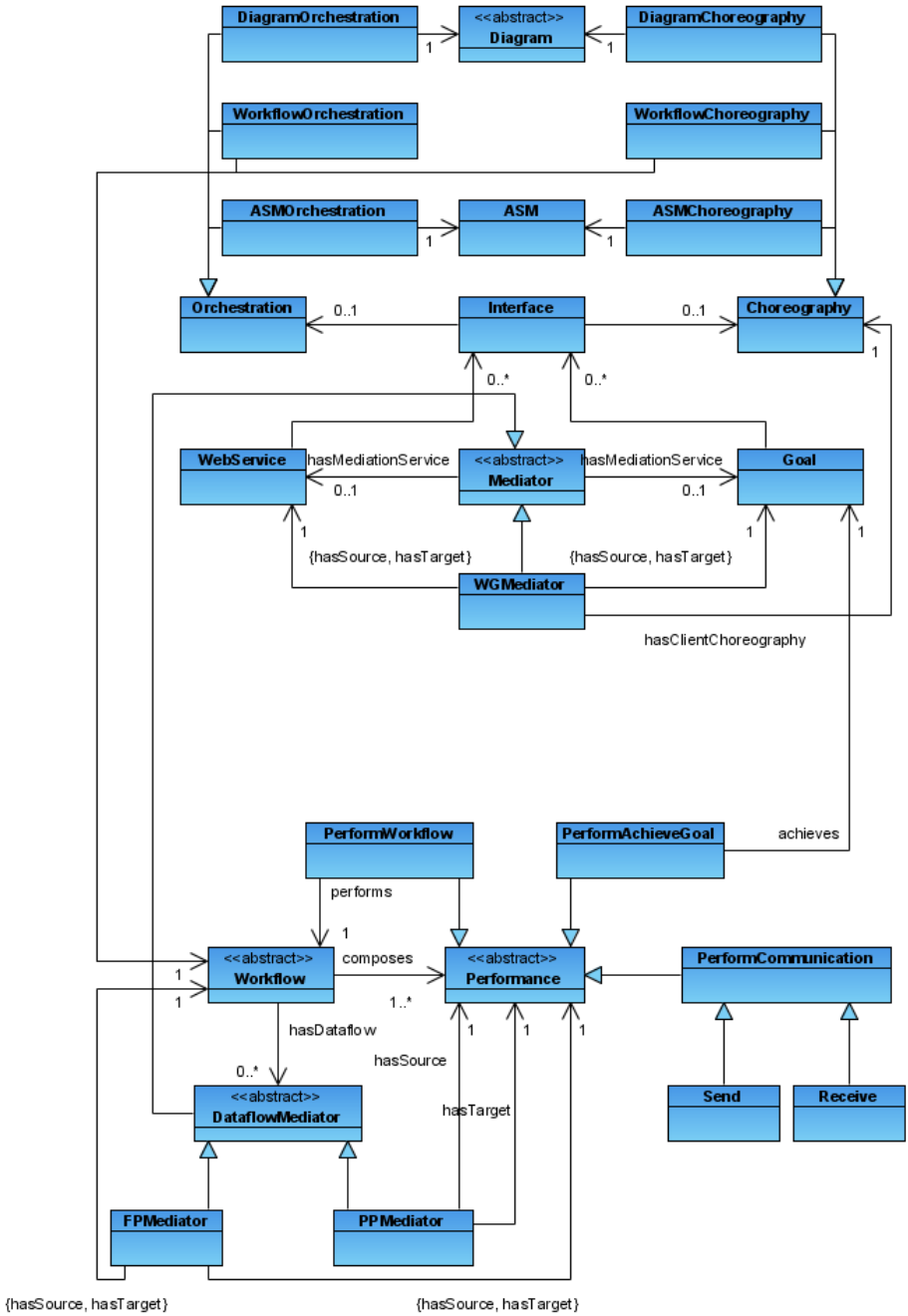


Fig. 2. Extended WSMO Meta-model

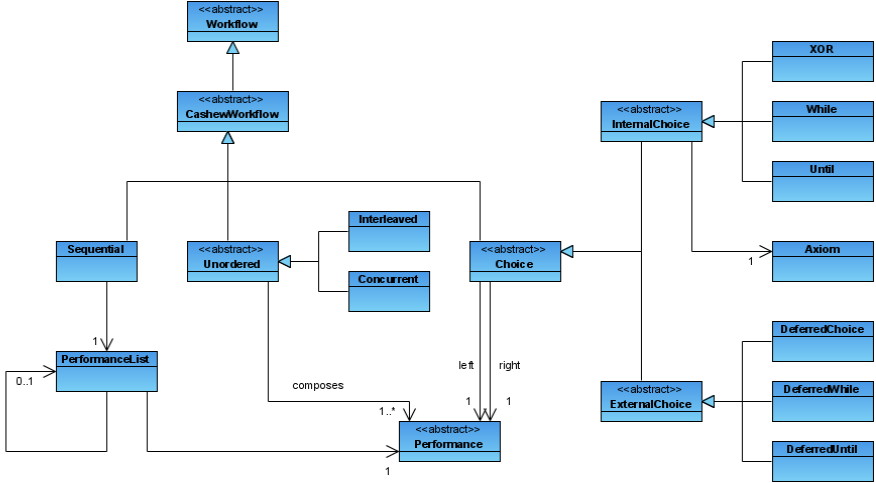


Fig. 3. Cashew Workflow Meta-model

The details of the Cashew workflow meta-model, extending the new concept *Workflow*, are diagrammed in Figure 3. The workflow operators are divided into three types: ‘Sequential’ depends on an ordered list of performances; ‘Concurrent’ — called ‘Split-Join’ in OWL-S — and Interleaved — called ‘Any-Order’ in OWL-S, and renamed as a shortened form of the Workflow Pattern ‘Interleaved Parallel Routing’ — both rely on an *unordered* set of performances; *choices* abstract over once-off choices and loops over exactly two performances.

The subset of operators to which ‘Internal Choice’ is a superconcept represent exactly those in OWL-S — with ‘If-Then-Else’ renamed after the Workflow Pattern ‘XOR’ — but substituting the WMSO concept ‘Axiom’ for the condition that the engine will evaluate to resolve the choice. In the case of ‘XOR’ the condition will only be evaluated once to chose between the left and right performance; in the case of ‘While’ and ‘Until’, after the left performance is evaluated, which will happen without evaluating the condition in the first instance with ‘Until’, the condition will be evaluated again.

In our previous semantics for OWL-S [22], we paid careful attention to the ‘Any-Order’ operator, elided in other semantics [1]. In the informal semantics published in the specification [10] it is stated that only one performance at a time will be executed, and that the performance to be executed at run-time will depend on availability of input data, since component performances may communicate to supply one another with data. This data-driven characteristic is in contrast to the control-driven ‘flow’ processes, due to WSFL [19], in BPEL [16].

In the spirit of this data-driven approach, since this happens to coincide with our own previous work [23], we offered an alternative semantics for ‘Choose-One’, where a non-deterministic choice would be made only between the ‘ready’ branches, *i.e.* those whose input has been provided. In the case that all branches

are ready, this is an equivalent non-deterministic choice. In the case that different outputs can be produced, *e.g.* by the invocation of an operation — not considered in OWL-S, but expected in WSMO — this allows the choice to be resolved externally between subsequent performances depending on the different messages. Furthermore, given the extension to explicit message receipts from the client, having extended the types of performance, this becomes the ‘classical’ deferred choice workflow pattern, which we therefore claim to generalise on, and name our operator after. We extend this ‘external choice’ to be able to decide also loops, as shown in the remaining concepts shown in Figure 3.

3 Representing Cashew in UML

The alliance with workflow patterns allows a standard mapping from most parts of Cashew directly into UML Activity Diagrams [27]. Rather than detail the whole translation here, we concentrate on the distinction between XOR and Deferred Choice discussed in the previous section.

For a XOR-type workflow between performances **V** and **W**, with axiom *a*, the resulting diagram fragment is as shown in Figure 4. A performance of this workflow would connect in control flow at the ‘decision node’ (the upper diamond), and connect it out at the ‘merge node’ (the lower diamond). It is an important part of the translation that each stage defines these two points uniquely, in order to be composable.

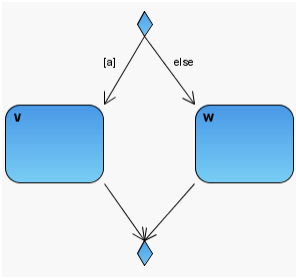


Fig. 4. XOR in UML

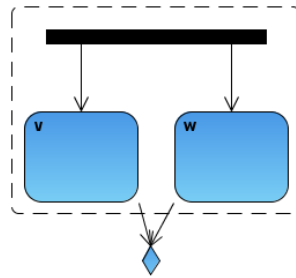


Fig. 5. Deferred Choice in UML

The Deferred Choice-type workflow between performances **V** and **W** is shown in Figure 5. Although this starts concurrently — the incoming control flow connects to the ‘split node’ (the horizontal bar) — there is within an ‘interrupting region’ (the dashed box) and each ‘interrupting edge’ preempts the other.

This can be contrasted with the representation of a performance of a concurrent workflow over performances **W1 .. Wn**, shown in Figure 6, where there is no interruptable region and the outgoing control flow resynchronises on a ‘join node’, rather than the ‘merge node’ in Figure 5, so every thread must complete.

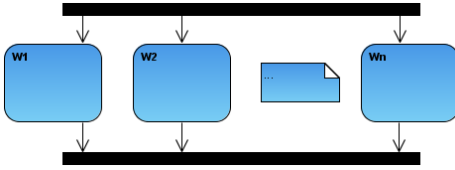


Fig. 6. Concurrent in UML

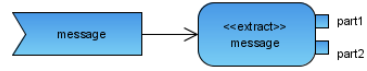


Fig. 7. Receive in UML



Fig. 8. Send in UML

In order to diagram the performance of sequential workflows, we simply create a control flow arrow between the representation of each successive performance, with the incoming control flow connecting to the first, and out-going to the last.

In order to define loops we allow the control flow to connect back to the ‘left’ performance via a ‘merge node’, after either a ‘decision node’ or an interruptible region in the form of Figures 4 and 5 respectively.

In order to easily define dataflow, each send and receive performance is associated with an ‘action’ with ‘pins’ to show the unpacked version of each message, as shown in Figures 8 and 7. In this way, dataflow can be represented by connecting pins together with edges that UML calls ‘object flow edges’. In fact, in WSMO terms, the dataflow edges represent mediators, as proposed in [3], and it is up to the mediators to specify the relationship between the message and the part needed, in the capability of their associated mediation goal or service, but this is a grammatical convenience.

4 Conclusions and Further Work

In this paper we have shown a three-level representation of behavioural models and how this is compatible with WSMO. We have shown how Cashew and the UML can be fitted into this model, allowing a relationship to be established with many existing communities. Finally, we have sketched how translation from a Cashew workflow model to an Activity Diagram model can be carried out.

The key advantages of the three-level model are the ability to deal with services with complex choreographies within orchestrations, and also to abstract from the details of sessions with these services in a goal-oriented fashion. The key advantages of Cashew are the ability to express the two kinds of choreography necessary to do this, including the distinction between internal and external choices which our example has shown is core to the WSMO notion of choreography, and the ability to communicate with the practitioners and tools of other communities, via the alliance with workflow patterns and UML diagrams.

Existing implementation involves an interpreter for Cashew in the IRS [4], and an on-going implementation of translation from Cashew to UML and a (partial) reverse-translation. Future work involves implementation of translations from

Cashew and Activity Diagrams to Abstract State Machines, and implementation of an orchestration engine based on Abstract State Machines in WSMX [18], the open-source reference implementation for WSMO.

Further on-going work involves the creation of an extended WSML grammar in which interfaces can be expressed in both Activity Diagrams and Cashew, as well as ASMs, and the proposal of further parts of the three-level model, with these concrete instances, to the WSMO/WSML Working Groups⁵.

Within the SUPER project⁶ the intention is to find a synergy between business process management and semantic web services. Early results in this project suggest that a three-level model, where an upper diagram-oriented representation is based on event-driven process chains (EPCs) [25], a middle level on a 'semanticised BPEL' and the bottom-layer on SWS-oriented representation may be useful. It is hoped that the three-level model here can be extended to express this in WSMO-compatible terms

Acknowledgements

This work is supported by the DIP project, an Integrated Project (no. FP6 - 507483) supported by the European Union's IST programme. We should like to particularly acknowledge the contribution of the DIP partner ILOG, specifically Laurent Henocque and Mathias Kleiner.

References

1. A. Ankolekar, F. Huch, and K. Sycara. Concurrent semantics for the web services specification language DAML-S. In *Proc. 5th Intl. Conf. on Coordination*, volume 2315 of *LNCS*, 2002.
2. E. Börger and R. Stärk. *Abstract State Machines*. Springer, 2003.
3. L. Cabral and J. Domingue. Mediation of semantic web services in IRS-III. In *Proc. Workshop on Mediation in Semantic Web Services (MEDIATE 2005)*, in conjunction with *ICSOC 2005*, 2005.
4. J. Domingue, L. Cabral, F. Hakimpour, D. Sell, and E. Motta. IRS-III: A platform and infrastructure for creating WSMO-based semantic web services. In *Proc. of the Workshop on WSMO Implementations (WIW 2004)*, volume ISSN 1613-0073. CEUR Workshop Proceedings, 2004.
5. J. Domingue, S. Galizia, and L. Cabral. Choreography in IRS-III: Coping with heterogeneous interaction patterns. In *Proc. 4th Intl. Semantic Web Conference (ISWC 2005)*, number 3729 in *LNCS*, 2005.
6. M. Dumas and A. H. M. ter Hofstede. UML Activity Diagrams as a workflow specification language. In *Proc. 4th Intl. Conf. on the Unified Modeling Language (UML)*, number 2185 in *LNCS*, 2001.
7. D. Roman *et al.* Orchestration in WSMO (working version). <http://www.wsmo.org/TR/d15/v0.1/>, January 2005.

⁵ <http://www.wsmo.org/>

⁶ <http://super.semanticweb.org/>

8. D. Roman *et al.* Web service modeling ontology WSMO v1.2. <http://www.wsmo.org/TR/d2/v1.2/>, April 2005.
9. D. Roman *et al.* Ontology-based choreography of wsmo services v0.3. <http://www.wsmo.org/TR/d14/v0.3/>, May 2006.
10. David Martin *et al.* OWL-S: Semantic markup for web services. <http://www.daml.org/services/owl-s/1.1/overview/>, 2004.
11. J. Kopecky *et al.* WSMO use case: Amazon e-commerce service v0.1. <http://www.wsmo.org/TR/d3.4/v0.1/>, December 2005.
12. M. Stollberg *et al.* DIP interface description ontology. <http://dip.semanticweb.org/documents/DIO-Annex-to-D3.4-and-D3.5.pdf>, January 2005. Annex to DIP Deliverables D3.4 and D3.5.
13. N. Kavantzias *et al.* Web services choreography description language v1.0. <http://www.w3.org/TR/ws-cdl-10/>, November 2005.
14. S. Bhiri *et al.* An orchestration and business process ontology. <http://dip.semanticweb.org/documents/D3.4.pdf>, January 2005. DIP Deliverable D3.4.
15. S. Galizia *et al.* An ontology for web service choreography. <http://dip.semanticweb.org/documents/D3-5.pdf>, January 2005. DIP Deliverable D3.5.
16. S. Thatte *et al.* Business process execution language for web services version 1.1. <ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf>, 2003.
17. Object Management Group. UML 1.4.2 specification. Technical Report ISO/IEC 19501, ISO, 2005.
18. A. Haller, E. Cimpian, A. Mocan, E. Oren, and C. Bussler. WSMX - a semantic service-oriented architecture. In *Proc. 4th Intl. Semantic Web Conference (ISWC 2005)*, number 3729 in LNCS, 2005.
19. F. Leymann. Web services flow language (WSFL 1.0). <http://www-3.ibm.com/software/solutions/webservices/pdf/WSFL.pdf>, 2001.
20. B. Norton. Experiences with OWL-S, directions for service composition: The Cashew position. In *OWL: Experiences and Directions Workshop (co-located with ESWC 2005)*, 2005. <http://www.mindswap.org/OWLWorkshop/sub23.pdf>.
21. B. Norton. Dataflow for orchestration in WSMO. <http://www.wsmo.org/TR/d15/d15.1>, July 2006.
22. B. Norton, S. Foster, and A. Hughes. A compositional semantics for OWL-S. In *Proc. 2nd Intl. Workshop on Web Services and Formal Methods (WS-FM 05)*, number 3670 in LNCS, Sept 2005.
23. B. Norton, G. Lüttgen, and M. Mendler. A compositional semantic theory for synchronous component-based design. In *14th Intl. Conference on Concurrency Theory (CONCUR '03)*, number 2761 in LNCS. Springer-Verlag, 2003.
24. M. Pistore, F. Barbon, P. Bertoli, D. Shaparau, and P. Traverso. Planning and Monitoring Web Service Composition. In *Proceedings of the Workshop on Planning and Scheduling for Web and Grid Services held in conjunction with ICAPS 2004, Whistler, British Columbia, Canada, June 3-7, 2004*.
25. W. M. P. van der Aalst. Formalization and verification of event-driven process chains. *Information & Software Technology*, 41(10):636–650, 1999.
26. W. M. P. van der Aalst, A. H. M ter Hofstede, B. Kiepuszewski, and A. P. Barros. Workflow patterns. *Distributed and Parallel Databases*, 14(3):5–51, June 2003.
27. P. Wohed, W. M. P. van der Aalst, M. Dumas, A. H. M. ter Hofstede, and N. Ruseell. Pattern-based analysis of UML activity diagrams. BETA Working Paper Series WP 129, Eindhoven University of Technology, 2005.

Learning from an Active Participation in the Battlefield: A New Web Service Human-Based Approach

Mauricio Paletta¹ and Pilar Herrero²

¹ Departamento de Ciencia y Tecnología. Universidad Nacional Experimental de Guayana.
Av. Atlántico. Ciudad Guayana. Venezuela
mpaletta@uneg.edu.ve

² Facultad de Informática. Universidad Politécnica de Madrid. Campus de Montegancedo
S/N. 28.660 Boadilla del Monte. Madrid. Spain
pherrero@fi.upm.es

Abstract. Real-time animation of virtual humans, capable of interacting realistically with others, requires a sophisticated architecture for the integration of different elements. A more flexible architecture for intelligent virtual agents emphasizing on the learning process should be designed to fulfil these requirements. In this paper we present an open and flexible architecture, IVAL, that has been designed to accomplish the requested necessities. IVAL is based on the Web Service principles, as well as on the fundamentals of the Open Agent Architecture (OAA) and it intends to accomplish determined objectives through the cooperation with other agents that inhabit the environment. One of the main purposes is to introduce a realistic learning process based on the interaction with the environment as similar as possible to humans been doing. This paper also presents a set of languages, based on the standard Extensible Markup Language (XML), that have been designed to get a more appropriate representation of all the required information elements.

Keywords: Intelligent virtual agent, learning, Web Service, XML.

1 Motivation and Related Work

A new emerging research line in the area of intelligent agents combines previous experiences in reasoning, planning and learning systems, with the experimental programming in the Virtual Reality Modeling Language (VRML). This has induced the development of Intelligent Virtual Agents (IVA), intelligent agents that can observe, decide and react in virtual environments.

Even though this field is being hardly demanded, presently there is not any architecture that could satisfy all groups of functionalities that users expect to get from an IVA in general. In this sense, there are two possible ways to implement the IVA: 1) define an own architecture, or 2) use an existing one. Both options require additional investigation efforts and could result in limitations or changes in the initial development approach.

In the same order of ideas, one of the processes that help in achieving the objectives previously indicated is the apprenticeship [1]. Learning plays a

fundamental role in many of the human activities since human learns from their experience, not only their achievements but also their errors [2]. In this way, if a specialized architecture in intelligent virtual agents needs to be defined, it is important to pay attention to the learning topic, which includes all an agent has to learn, how it should learn it and how knowledge should be handled.

There have been several works that have exploited the idea of open and flexible architectures. Martin et al [3] explain the structure and elements for the construction of systems based on agents using OAA¹ (Open Agent Architecture), in which various agents help each other under the premise of requiring and/or providing services based on their capabilities. In this approach, each agent that participates in the system defines and publishes a group of capabilities through ICL (Inter-agent Communication Language), also defined in OAA. In addition, one facilitator maintains a base of knowledge that registers the capabilities of the group of agents and uses this knowledge to help service applicants and suppliers.

In [4] the authors consider an open architecture for intelligent agents. They refer to the incorporation of new agents when new appropriate methods to the area in which their architecture was focused are developed. In the same way, De Antonio et al [5] show an architecture for the development of intelligent agents based on a set of cooperative software agents. The idea of using agents to structure the architecture of the intelligent agent is exploited here.

On the other hand, there have been also other important works that have correlated a conceptual relation between the theories of web services and software agents. In order to describe the available services for one agent, specifically BDI agents, Dickinson et al [6] identify the importance for each agent to have meta-data for this purpose.

Motivated on this previous works, IVAL, an open and flexible IVA architecture is presented in this paper, describing some of the elements that conform this architecture as well as giving a special consideration to the learning topic. IVAL is one of the first results of the research work that is being carried out at the Universidad Nacional Experimental de Guayana (Venezuela) in close collaboration with the Universidad Politécnica de Madrid (Spain). In order to support the characteristic of an open architecture, XML² (Extensible Markup Language), due to its simplicity and flexibility, is used as the base for the representation of all the information items necessary to cover the IVA functionalities.

2 Designing the Open and Flexible Multi Agent System

The designed architecture, named IVAL (Open and Flexible Architecture based on the IVA Learning Process) has been inspired in the concept of SOA (Service-Oriented Architecture) and the fundamentals of OAA. SOA is extensively used at present for the design of development models of Internet systems, being the Web Services technology³ one of the most important. The Web services and the software agents share a motivation in searching mayor flexibility and adaptability for the information

¹ <http://www.ai.sri.com/oa/>

² <http://www.w3.org/XML/>

³ <http://www.w3.org/2002/ws/>

systems; that is why it is natural to consider a conceptual relation between these two technologies [6].

On the other hand, OAA defines a framework for the construction of distributed software systems. Its theoretical bases, that describe the structure and necessary elements to configure a cooperation environment among agents, has been useful in contributing with ideas related to the concept of flexible and open architectures, in which agents can be easily incorporated into, or unincorporated from, the system.

In this sense, IVAL is an architecture based on a group of agents that cooperate together to achieve the specific objectives of the IVA. Each one of these agents is specialized in a particular service associated with some abilities: deliberate, react, socialize, interact with the environment, learn, and others. These agents are divided into one Agent of Control (AC) and several Agents of Service (AS).

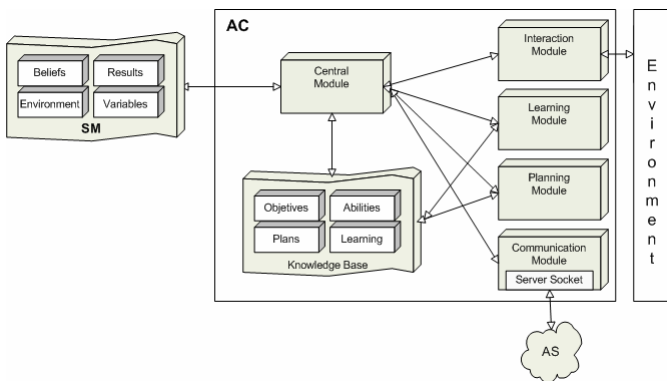


Fig. 1. Internal structure of the AC

The AC is the main piece of the IVAL architecture. Its main function is to structure itself with all the abilities that, as a group, offer not only the same AC but also all the AS with which it interacts with. This allows this IVA to be seen, from outside, by the remainder IVA's that surround it, like an agent equipped with all these abilities. Additionally, the AC is the key element to make this architecture extensible. In order to comply with these functions, the AC will have (see Fig. 1):

- A knowledge base responsible of structuring: the IVA objectives; its planning; the current abilities provided by the AS with which is interacting, and the knowledge learned through the corresponding AS.
- The following functional modules:
 - Communication module: it takes care to handle not only the socket of connection but also the communication protocol with the AS.
 - Planning module: it contains the algorithm that allows the IVA to execute its action plans of for the objectives achievement.
 - Learning module: it contains the set of learning basic algorithms of the IVA; it integrates the learning with the corresponding AS for this ability.
 - Interaction module: it handles the basic interaction of the IVA with the environment.

- Central module: it synchronizes the execution of the remainder modules and allows the interaction of the SM with the remainder agents.

The AC functions can be summarized as follow: have a basic interaction with the environment, establish the final objectives of the IVA and prepare the plan to satisfy these objectives, communicate themselves with the AS, verify that the AS perform their tasks, synthesize the information received from the AS, and, manage learning with the specialized AS in this ability.

The UML collaboration diagram illustrated in Fig. 2 can be reviewed in order to describe in more detail the relation among the elements that conform the structure of the AC. The labeling links do not represent a specific order and are interpreted as follow: 1- receiving incentive from the environment; 2- updating the SM; 3- learning; 4- updating knowledge; 5- communicating to AS (with learning abilities); 6- receiving answer from the AS; 7- obtaining knowledge related to the plans and objectives; 8- updating results; 9- reporting results; 10- arranging the performance, and 11- acting.

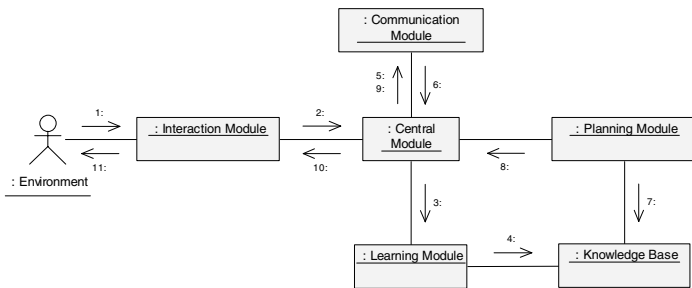


Fig. 2. Collaboration diagram among AC elements

The main function of the AS, on the other hand, is to cover the ability that has been conferred to them, in addition to keep communicating their results with the AC. They can be elaborated following architectures of agents already known, specialized in satisfying specific abilities. Based on the Belief-Desire-Intention (BDI) architecture [7] an approach of an AS has certain similarity with the AC with some particular differences. It has a knowledge base with the following information elements: the objectives of the agent (from the service), the plans for the achievement of these objectives, specialized knowledge required according to the agent ability, and the variables and internal states. To incorporate the AS into the IVAL architecture, the AS has to implement the communication mechanisms with the AC, as a minimum.

The AS can be added / removed from the agent system originating an increase / decrease of the IVA abilities. Each agent handles its own knowledge base and / or partial memory necessary to comply with its own specific objectives. These AS do not have direct interaction with the environment unless they are related to specialized services associated with typical functionalities of interface with the environment: sound management, images processing, natural language processing, etc. The same service can be given for various agents, which permits to reinforce the IVA ability associated with this service. Any possible inconsistencies shall be resolve by the AC.

The way in which an AS is incorporated into the IVA structure follows the same idea considered in the Web Services with the publication of the services. As in the case of a WSDL (Web Service Description Language) document, that describes the content of a Web Service, IVAL-SADL (Service Agent Description Language for IVAL), based on the XML language, has been developed in this work, as a subgroup of the labels or descriptive elements defined by DAML⁴ (DARPA Agent Markup Language) and AML⁵ (Agent Markup Language).

The IVAL-SADL of an agent contains three levels of information: basic information related to the implementation, such as the version and author, in the first level; the set of abilities that the agent implements, in the second level; and the set of service or protocol commands to which the agent can respond in the third level.

The communication between the AS and AC is carried out by flows of information in XML format, in the same way as SOAP (Simple Object Access Protocol) is used in Web Service systems. A server/client TCP/IP communication channel is opened between an AS and the AC via sockets. The AC is listening continuously the channel to allow the AS to be connected. The first thing AS do when it is connected is to send the corresponding IVAL-SADL document that describes its abilities. The AC updates its knowledge base with the list of the current abilities of the IVA.

In order to achieve this, IVAL-VACL (Virtual Agent Communication Language for IVAL) language has been designed, based on the defined standard in the XML version of FIPA-ACL (Foundation for Intelligent Physical Agents-Agent Communication Language). Some labels described in the works of Choi et al [8] and Makatchev et al [9] were also considered.

This new architecture contains a memory area shared among all the agents (SM – Shared Memory), which it is located in a space, preferably in the same location where the AC resides in. All agents shall know the location of this memory to be able to access to it. This SM is formed by XML based documents that represent and store the description of the environment, the beliefs of the IVA, the conclusions and results of the agents called services, and the variables and internal states.

As a complement of the IVAL architecture, the following section describes the way in which the learning process is carried out.

3 The IVAL Learning Process

As it has been mentioned previously, the IVAL basic architecture is equipped with learning elementary abilities, thanks to the corresponding module in the AC structure (Fig. 1). One of the fundamental functions of this module is to maintain the component training of the AC knowledge base. Since learning is one of the abilities or specializations that can be associated with an AS, the other function of this module is related to the coordination of possible AS with learning abilities existing in the architecture. The AS defined to cover this ability are informed, through the AC, about everything that happens in the SM of the IVA as well as everything that happens in the rest of the environment when the remainder agents complete any task or satisfy any of their objectives.

⁴ <http://www.daml.org/>

⁵ http://www.bingooo.com/images/BINGOOO_syntax_1_1-engl.pdf

The knowledge base learned from the IVA is initially empty. As the IVA is acting based on received stimuli, the relation between what it is saw, heard, touched, smelled or tasted and the given answer is represented and stored. Any stimulus consists of a pair <data, type> that represents the information coming from the environment whether it is by the interaction module of the AC or by any AS with the environment interaction ability. The reaction of the IVA consists in series of executed services and their corresponding results. In this sense, a rule of knowledge between the stimulus (the condition) and the set of services after the stimulus reception (the consequence) can be written. In order to represent this experience, a XML based document is writing using the new language IVAL-KBEL (Knowledge Base Experience Language for IVAL).

For each new stimulus, the AC looks for an adequate answer, whether it is with its internal modules or with the support of the AS. Simultaneously, the AC consults to the learning module if there is any experience with the present state or stimulus. In this case, all the services associated with the stimulus in the knowledge rule stored into the knowledge base, and with the adequate results, are called in the same way they are called the first time when the same stimulus was received. If another different service is called by the user, the information is considered to reinforce or give feedback of the knowledge according to what it was learned. The relationship between one stimulus and the executed services, that is to say the knowledge rule, finishes when a new stimulus comes.

The IVAL knowledge base is then a set of knowledge rules based on the relationship <stimulus, action>, being the action represented by the executed services in the IVAL context. So the stimulus reception as the execution of the services depends on the AS connected or integrated to the IVAL in the actual moment. Due to the open and extensible characteristics of IVAL, it is not possible to repeat the same scenario in response to the same stimulus. Without considering what AS is connected or not to the IVAL, all the experience is considered and stored in the knowledge base.

4 Implementation and Evaluation

This section describes the way in which IVAL was implemented and evaluated in a scenario of interaction with the environment. In this sense, a C# class library was developed in order to implement the concepts related to IVAL. This library contains classes for the definition of each AC module, the SM, and each one of the necessary XML documents. Fig. 3 shows the UML class diagram associated with the C# library.

For evaluation purposes, each IVA has been developed to not only interact with the environment through, at least, one of its five senses, but also learn from its acting. Each sense's capability to interact with its surroundings is carried out by corresponding specialized AS services. In this sense, the ASSound agent has been implemented, based on the SADL document (Fig.4), to manage the sound (listening and talking). In the same way, the ASTouch, ASSmell, ASSee and ASTaste agents have been implemented for the corresponding senses and skills. Another agent, ASFace, has also been implemented to represent a set of face's expressions like happiness, sadness, fear and timidity.

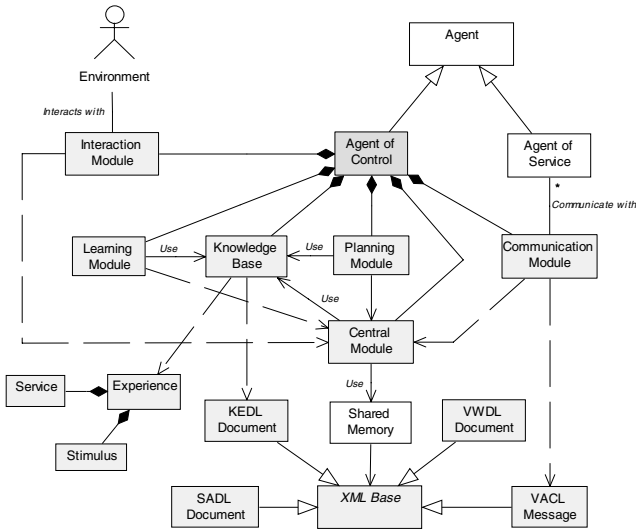


Fig. 3. Class diagram of the IVAL C# library

```
<?xml version="1.0" encoding="ISO-8859-15"?>
<IVAL-SADL:Agent name="ASSound" type="environment"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="IVAL-SADL.xsd" xmlns:IVAL-SADL="http://www.Ival.org">
  <IVAL-SADL:Basic>
    <IVAL-SADL:Version>1.0</IVAL-SADL:Version>
    <IVAL-SADL:Author>Mauricio Paletta</IVAL-SADL:Author>
    <IVAL-SADL:Author_url></IVAL-SADL:Author_url>
    <IVAL-SADL:Author_email>mpaletta@uneg.edu.ve</IVAL-SADL:Author_email>
  </IVAL-SADL:Basic>
  <IVAL-SADL:Abilities quantity="1">
    <IVAL-SADL:Ability>Handle the sound</IVAL-SADL:Ability>
  </IVAL-SADL:Abilities>
  <IVAL-SADL:Services quantity="3">
    <IVAL-SADL:Service name="BeginToListen">
      <IVAL-SADL:Result type="bool"/>
    </IVAL-SADL:Service>
    <IVAL-SADL:Service name="EndToListen">
      <IVAL-SADL:Result type="none"/>
    </IVAL-SADL:Service>
    <IVAL-SADL:Service name="ToTalk">
      <IVAL-SADL:Parameter name="Phrase" type="string"/>
      <IVAL-SADL:Result type="bool"/>
    </IVAL-SADL:Service>
  </IVAL-SADL:Services>
</IVAL-SADL:Agent>
```

Fig. 4. The IVAL-SADL document of the ASSound agent

The evaluation was carried out with an IVA endowed with the IVAL internal architecture and the incorporation of six AS. The IVAL basic architecture and the AS are independent applications with a simple interface that simulates the interaction with the environment.

Different scenarios of interaction between the IVA and the environment (composed by the user and others IVA), as well as the corresponding services were raised for evaluation purposes. Among all these possible scenarios, one of them could be, for example, an application designed for military training in which the soldiers could be

trained to be faced to dangerous situations. This could be represented in the IVAL context as follows:

1. Thinking in a situation in which a soldier, who is alone in a mission, listens a noise close to a thicket, and decides to inspect what the source of this noise is; this could be represented in the IVAL as the ASSound agent informing to the AC using an IVAL-VACL message about the stimulus <“noise”, “listened”>. This is possible because the ASSound is an AS with the ability to interact with the environment.
2. The soldier faces another soldier and he is not sure if it is ally or enemy. This is, the AC receives the stimulus and looks for experience associated with this stimulus. As the knowledge base is initially empty, a logger message of “I don’t know what to do” is given. As a consequence, IVAL user should intervene and the IVAL learning process learns what to do when the same stimulus is received again in the future. Steps 3 to 7 summarize this process.
3. As the situation is not secure and the soldier doesn't feel save, he decides to keep quiet, using his visual perceptual stimulus as a part of his strategy to go unnoticed in this unfortunate situation; therefore, the AC interaction calls to the “ToSee()” service associated to the ASSee agent instead of using the ASSound to talk . To do this, a corresponding IVAL-VACL message (Fig. 5) is sent from the AC to this AS.
4. Due the fact that the soldier could have or not the ability to capture visual stimulus, the possession of this ability is represented by the ASSee agent. In this sense, when the ASSee executes the service received in the message, it sends to the AC another similar message with the corresponding result that indicates whether was or not possible to execute the service. It is important to mention that once the AC receives a message from any AS with the results associated with the services previously called, the AC updates the shared memory into the corresponding “Results” component, and the knowledge rule associated with the actual stimulus.
5. If there are not problems with the soldier’s vision and since he is close to a thicket, the ASSee agent sends to the AC the stimulus <“thicket”, “seen”> as a consequence of the “ToSee()” execution.
6. At this point, it is time the soldier learns what have happened until now after receiving the previously stimulus and before deciding what to do with this new stimulus. In this case, the AC receives the new stimulus and, before repeating the step 1 again, it updates the knowledge base with the knowledge rule associated with the previously stimulus. An example of how the knowledge base is stored at this moment, using the IVAL-KBEL language, is illustrated in Fig. 6.
7. Since the IVA soldier doesn’t know what to do with this new stimulus, the IVAL user intervenes again indicating him to go to the thicket and see again. The IVAL learning process continues in the next steps.
8. Using his visual perceptual stimulus again, the soldier sees an unknown soldier hidden in the thicket, so, the next stimulus the AC received from the ASSee agent is <“unknown soldier”, “seen”>.
9. Based on the fact that, in the real situation, reviewing the face expression of the other person could help soldier to take the appropriate decision, in response to the

stimulus described previously, the AC is then commanded to see the unknown soldier's face expression via the ASSee agent.

10. Imaging that the person face expresses fear, the soldier decides then to be friendly and ask him/her to remain calm. In this case, the received stimulus is <"fear expression", "seen">, and the reaction to this stimulus is to call the ASSound service "ToTalk("you remain calm")", the ASFace service "ToSet("friendly")" and the ASSee service "ToSee()" again.
11. After that, the soldier sees that the face expression of the unknown soldier changes to a friendly expression allowing him to conclude that the situation is now under control. In this case, the stimulus received by the AC is now <"friendly expression", "seen"> for which the IVA is intended to react suitably.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<IVAL-VACL:Message sender="IVAL Military Training"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="IVAL-VACL.xsd" xmlns:IVAL-VACL="http://www.Ival.org">
  <IVAL-VACL:Service>
    <IVAL-VACL:Name>ToSee</IVAL-VACL:Name>
    <IVAL-VACL:Result>none</IVAL-VACL:Result>
  </IVAL-VACL:Service>
</IVAL-VACL:Message>
```

Fig. 5. An example of one IVAL-VACL message sent from the AC to ASSound agent

```
<?xml version="1.0"?>
<IVAL-KBEL:Knowledge xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="IVAL-KBEL.xsd" xmlns:IVAL-KBEL="http://www.Ival.org">
  <IVAL-KBEL:Experience>
    <IVAL-KBEL:Stimulus source="listened">
      <IVAL-KBEL:Data>Noise</IVAL-KBEL:Data>
    </IVAL-KBEL:Stimulus>
    <IVAL-KBEL:Service name="ToSee">
      <IVAL-KBEL:Result>true</IVAL-KBEL:Result>
    </IVAL-KBEL:Service>
  </IVAL-KBEL:Experience>
</IVAL-KBEL:Knowledge>
```

Fig. 6. IVAL knowledge base example

5 Conclusions and Ongoing Research Work

Through this work, it was empathized the importance of using the concepts of SOA and OAA to define the IVAL specifications and how the relation between agent and service has helped to facilitate these concepts.

Even though IVAL was originally designed for an IVA, it could be applied to other kind of software/intelligent agents such as a mobile agent or an agent advisor, since it basically focuses on the interaction with the environment and the learning process.

As the communication between the service agents and the control agent is through the TCP/IP protocol, the service agents and the control agent could be located in different physical spaces, and the cooperation between them could still be possible.

A set of XML based languages have been defined in order to represent and manage the information associated with the environment, as well as the description of the service agents, the communication between the service agents and the control agent, and the agent control related knowledge.

Based on the evaluation of the model, it was found that specialized agents could be integrated to the IVA systems in order to manage specific abilities, in the same way as humans could have or not particular abilities. On the other hand, the evaluation also demonstrated that IVAl learns from the experience of the interaction with the environment starting with an empty knowledge base, in the same way as humans start learning since they are born.

Once the library and the correct integration between the AC and the AS has been evaluated in simple scenarios, the next step will be to integrate this library in more complex systems and virtual environments in order to produce real applications related to IVA's. The knowledge rule will also be modified in order to incorporate an acceptance value of any service associated with the stimulus. This will permit IVA to learn what it is good to do or not in response to any received stimulus.

References

1. Herrero P., de Antonio A.: A Human Based Perception Model for Cooperative Intelligent Virtual Agents. Lecture Notes in Computer Science, Vol. 2519, 195-212, Springer, ISBN: 3-540-00106-9. 10th International Conference on Cooperative Information Systems (CoopIS'2002), Irvine, California, USA, (2002).
2. Buczak A. L., Cooper D. G., Hofmann M. O.: The Evolutionary Platform for Agent Learning. Artificial Neural Networks in Engineering, ANNIE, St. Louis, MO, USA. (2003).
3. Martin D. L., Cheyer A. J., Moran B. D.: The Open Agent Architecture: A Framework for Building Distributed Software Systems. Applied Artificial Intelligence, Vol. 13, Nos. 1-2, 21-128. (1999).
4. Gurer D., Lakshminarayan V., Sastry A.: An Intelligent-Agent-Based Architecture for the Management of Heterogeneous Networks. Proc. Ninth Annual IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, Delaware, USA. (1998).
5. De Antonio A., Ramírez J., Méndez G.: An Agent-Based Architecture for Virtual Environments for Training. In Developing Future Interactive Systems, Chapter VIII, Idea Group Publishing, ISBN 1591404118. (2005).
6. Dickinson I., Wooldridge M.: HP Labs 2005 Technical Reports. (2005).
7. Rao A. S., Georgeff M. P.: Modeling Rational Agents within a BDI-Architecture. Proc. Second International Conference on Principles of Knowledge Representation and Reasoning, San Mateo, CA, USA. (1991).
8. Choi J., Kim Y., Kang S.: An Agent Communication Language Using XML and RDF. (2001).
9. Makatchev M., Tso S. K.: Human-Robot Interface Using Agents Communicating in an XML-Based Markup Language. Proc. IEEE International Workshop on Robot and Human Interactive Communication, Osaka, Japan. (2000).

A Collaborative Awareness Specification to Cover Load Balancing Delivery in CSCW Grid Applications

Pilar Herrero¹, José Luis Bosque², Manuel Salvadores^{1,3}, and María S. Pérez¹

¹ Facultad de Informática.
Universidad Politécnica de Madrid.
Madrid. Spain
{pherrero, mperez}@fi.upm.es

² ESCET
Universidad Rey Juan Carlos, Spain
joseluis.bosque@urjc.es

³ Imbert Management Consulting Group
Madrid Spain
mso@imcs.es

Abstract. In this paper, we present a new extension and reinterpretation of one of the most successful models of awareness in Computer Supported Cooperative Work (CSCW), called the Spatial Model of Interaction (SMI), which manage awareness of interaction through a set of key concepts, to manage task delivery in collaborative distributed systems. This model, called AMBLE (Awareness Model for Balancing the Load in Collaborative Grid Environments), also applies some theoretical principles of multi-agents systems to create a collaborative environment that can be able to provide an autonomous, efficient and independent management of resources available in a Grid. This model has been implemented using web services and some experimental results carried out over and real and heterogeneous grid are presented with the end of emphasizing the performance speedup of the system using the AMBLE model.

1 Introduction

Grid computing intends to share heterogeneous resources in dynamic environments. The complexity of achieving this goal is caused by several factors, being the existence of different virtual organizations, the dynamism of the underlying architecture and the heterogeneity of the involved resources some of the most challenging aspects.

With the aim of providing better capabilities on a grid, it is essential to use a resource manager, which will take the decision about the allocation of processes to the resources of the system. The resource management includes other tasks, such as resources discovery, resources registration and monitoring. As expected results, the resource manager should achieve load balancing within the grid. Equilibrating the amount of work assigned to each node in a grid is a complex problem, even more than for other kinds of parallel systems. In [8], authors apply reinforcement learning to adaptive load balancing for allocating resources in a grid in an efficient way. In [4], a communication-based load balancing algorithm, named Comet, is shown.

On the other hand the multi-agent systems offer promising features to resource managers. The reactivity, proactivity and autonomy, as essential properties of agents, can help in the complex task of managing resources in dynamic and changing environments. Additionally, the cooperation among agents allows load balancing mechanisms to be performed and efficiently deployed on a grid. In [3], an agent-based grid management infrastructure together with a task scheduler is performed for local grid load balancing. Four different negotiation models are studied in [11] for agent-based load balancing and grid computing. The interaction between grids and agents challenges has been clearly defined by Foster et al in [7]. As example of the successful combination of grid and agents, a real grid system has been built by means of mobile agent technology, SMAGrid, Strong-Mobile Agent-Based Grid [12].

In this paper, we present a new extension and reinterpretation of the Spatial Model of Interaction (SMI), an abstract awareness model designed to manage awareness of interaction, in cooperative applications. A new reinterpretation of this model, and its key concepts, in the context of an asynchronous collaboration in grid environments is developed in this paper. This reinterpretation, open and flexible enough, merges all the OGSA [6] features with theoretical principles of multi-agents systems, to create a collaborative and cooperative grid environment. Following one of the main OGSA characteristics, the use of open, standard and public interfaces, we have implemented AMBLE as a web service specification, WS-AMBLE. This specification provides an open interface having the ability of managing different levels of awareness, allowing different Virtual Organizations to share computational resources based on open protocols and interfaces. As far as we know, none of the last WS specifications offers functionalities useful enough as to create awareness models that could be able to manage task balancing delivery in collaborative grid environments.

2 AMBLE: Reinterpreting the Key Awareness Concepts

The Spatial Model of Interaction, defined for application to any Computer Supported Cooperative Work (CSCW) system where a spatial metric can be identified [2]. The model itself defines some key concepts such as:

- * *Aura*: Sub-space which effectively bounds the presence of an object within a given medium and which acts as an enabler of potential interaction [5].
- * *Focus*: It delimits the observing object's interest.
- * *Nimbus*: It represents the observed object's projection
- * *Awareness*: It quantifies the degree, nature or quality of interaction between two objects. For a simple discrete model of focus and nimbus, there are three possible classifications of awareness values when two objects are negotiating [9]: *Full awareness*, *Peripheral awareness* and *No awareness*.

Let's consider a system containing a set of nodes $\{n_i\}$ and a task T that requires a set of processes to be solved in the system. Each of these processes necessitates some specific requirements, being r_i the set of requirements associated to the process p_i , and therefore each of the processes will be identified by the tuple (p_i, r_i) and T could be described as $T = \sum_i \{(p_i, r_i)\}$. The AMBLE model, proposes an awareness infrastructure

based on these concepts capable of managing the load management of grid environments. This model reinterprets the SMI key concepts as follow:

Focus: It can be interpreted as the subset of the space on which the user has focused his attention with the aim of interacting with. Regarding tasks, and given a node n_i in the system requiring the execution of a given task (T), the focus of this node would contain, at least, the subset of nodes that are composing the Virtual Organization (VO) in which this node is involved

Nimbus: It is defined as a tuple ($Nimbus=(NimbusState ,NimbusSpace)$) containing information about: (a) the load of the system in a given time (*NimbusState*); (b) the subset of the space in which a given node projects its presence (*NimbusSpace*). As for the *NimbusState*, this concept will depend on the processor characteristics as well as on the load of the system in a given time (see section 4).

Awareness of Interaction (AwareInt): This concept will quantify the degree, nature or quality of asynchronous interaction between distributed resources. Following the awareness classification introduced by Greenhalgh in [9], this awareness could be *Full*, *Peripheral* or *Null*.

$$AwareInt(n_i, n_j) = Full \quad n_j \in Focus(\{n_i\}) \wedge n_i \in Nimbus(n_j)$$

Peripheral aware of interaction if

$$n_j \in Focus(\{n_i\}) \wedge n_i \notin Nimbus(n_j)$$

or

$$AwareInt(n_i, n_j) = Peripheral \quad n_j \notin Focus(\{n_i\}) \wedge n_i \in Nimbus(n_j)$$

The AMBLE model is more than a reinterpretation of the SMI, it extends the SMI to introduce some new concepts such as:

Interactive Pool: This function returns the set of nodes $\{n_j\}$ interacting with the n_i node in a given moment. Given a System and given a task T to be executed in the node n_i

Task Resolution: This function determines if there is a service (s_i) in the node n_i , being $NimbusState(n_i) \neq Null$, such that could be useful to execute the task T (or at least one of its processes).

$$n_i = \sum_i \{s_i\} \quad Task Resolution: \quad Node \times Task \rightarrow \quad Task$$

$$n_i \times T \rightarrow \{(p_i, s)\}$$

Where s is the “score” to execute p_i in n_i node, being its value within the range $[0, \infty)$: 0 if the node n_i fulfils the all the minimum requirements to execute the process p_i ; the higher is the surplus over these requirements and it will be the value of this score.

Collaborative Organization: This function will take into account the set of nodes determined by the *Interactive Pool* function and will return those nodes of the System in which it is more suitable to execute the task T. This selection will be made by means of the *TaskResolution* function.

3 Load Balancing Algorithm in AMBLE

In this section we will introduce the load balancing algorithm as it has been introduced in the AMBLE awareness model, and how it will be applied to our distributed and collaborative multi-agent architecture in grid environments

State Measurement Rule: This local rule will be in charge of getting information about the computational capabilities of the node in the system. This information, quantified by a load index, provides aware of the *NimbusState* of the node. This dynamic index should be periodically and frequently measured, and should be a good estimation of a node computing capabilities. The choice of a load index has a huge impact on load balancing efficiency [10]. In this paper the load index is evaluated based on two parameters:

- *Node computational power (P)*, which depends on the node computational architecture, and takes into account CPU, memory and I/O features. It is a static parameter.
- *The CPU assignment* which is defined as the percentage of CPU time that would be available to a newly created task, on a specific node. It will be working as a dynamic parameter.

The *NimbusState* of the node will be determined by the load index and it will depend on the node capacity at a given time. This state determines if the node could execute more (local or remotes) tasks. Its possible values would be:

- * *Maximum:* The load index is low and therefore this infrautilized node will execute all the local tasks, accepting all new remote execution requests coming from other nodes.
- * *Medium:* The load index has an intermediate value and therefore the node will execute all the local tasks, interfering in load balancing operations only if there are not other nodes whose *NimbusState* would be *Maximum* in the system.
- * *Null:* The load index has a high value and therefore the node is overload.

Information Exchange Rule: This policy should keep the information coherence of the global state of the system, without overloading the network with an excessive number of unnecessary messages. An optimum information exchange rule for the AMBLE model should be based on events [1]. This rule only collects information when a change in the *NiImbus* (in the *NimbusState* or in the *NimbusSpace*) of the nodes is made. The node that has modified its *nimbus* will be in charge of notifying this modification to the rest of the nodes in the system, avoiding thus synchronisation points. The information that every node has about the *NimbusState* of the rest of the nodes is updated while the node receives information messages from the others.

Load Balancing Operation: Once the node has made the decision of starting a new load balancing operation, this operation will be divided in another three different rules, presented in the following sections.

Localization Rule: It has to determine which nodes are involved in the CollaborativeOrganization of the node n_i . In order to make it possible, firstly, the AMBLE model will need to determine the awareness of interaction of this node with those nodes inside its focus. Those nodes whose awareness of interaction with n_i was

Full will be part of the Interactive Pool of n_i to solve the task T , and from that pre-selection the TaskResolution method will determine those nodes that are suitable to solve efficiently the task in the environment.

Selection and Distribution Rule: This algorithm joins selection and distribution rules because it determines which nodes will be in charge of executing each of the processes making up the T task. The proposed algorithm takes into account the NimbusState of each of the nodes as well as the TaskResolution to solve any of the T 's processes. The goal of this algorithm is to find the more equilibrate assignment of processes to computational nodes, based on a set of heuristics. Firstly, a complete distribution is made taking into account all the processes making up the T task as well as all the computational nodes implicated in the CollaborativeOrganization. If, in this first turn, it would be possible to assign all the process the algorithm would have finished. Otherwise, it would be necessary to calculate, again, the NimbusState of the nodes belonging to the CollaborativeOrganization, repeating the complete process again.

4 AMBLE Evaluation Architecture

As it is possible to appreciate in the Figure 1 the middleware architecture of load balancing model has been separated in three different parts:

- **SMI-Engine (Spatial Model of Interaction Engine):** This is the main core of the architecture and contains those components that implement all the logic of the SMI and the load balancing algorithms explained in section 6. This engine is made up by the following modules:
 - Benchmark Agent: This agent based on the Linpack benchmark [16], maintains a performance measure of the node which could be evaluated from the Load Balancer.
 - Global State Agent: This agent compiles all the information related to the two main concepts of the AMBLE, providing information about those nodes that are available in the system.
 - Load Balancer Agent: This module implements the logic for the load-balancing operation and makes the final decision of which node will execute each process.
 - Execution framework: This interface contains the modules dependent on the operating system (OS) to access to the process management APIs.
 - SO Native Process Management: It depends on the OS and uses those functionalities that the APIs of this OS offer to supply the to process execution.
 - SO Native Monitoring Module: This module also depends on the OS and uses those functionalities that the APIs of this OS offer to monitor the state of the computer.
- **AMBLE-Service:** Web service [13, 14] interface that provides those methods necessary to establish the communication between nodes through SOAP [15] messages. The operations deployed in this service are:
 - registerVisibility: When a node detects a new resource inside its focus, it invokes this operation. Moreover, if the observer node is also inside the

observed NimbusSpace, it would be included in the awareness of interaction record with a value equal to Full.

- **nimbusChangeCallback**: This operation receives the changes that a specific node has on its NimbusState.
- **requestTask**: This method is invoked by a client requiring the execution of the T task composed by n processes.
- **taskResolution**: This method is invoked by a node requiring “offers/scores” for the processes associated to a specific task.
- **performTask**: This method is invoked to order the process execution once the process has been assigned to a particular node.
- **monitorExec**: This operation is used to monitor the state of execution of a process in an identifiable node.

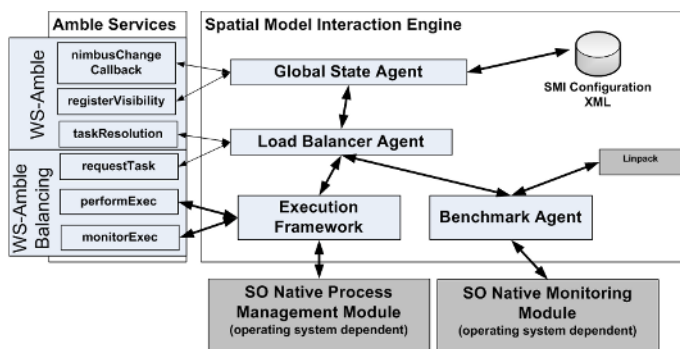


Fig. 1. The Load Balancing Model: Middleware Architecture

In Figure 2 it is possible to appreciate the sequence of operations unleashed inside a node after having received the registerVisibility message from another one:

1. The first step is to corroborate if the node that invokes the operation registerVisibility is inside its NimbusSpace.
2. If the remote node is inside its NimbusSpace, (enableReg=true), the remote node will add it to its awareness of interaction record with a value equal to Full, and the origin node will return its NimbusState. This execution branch determines that both nodes could collaborate, if needed, to carry out a load-balancing operation.
3. However, if the remote node is not inside its NimbusSpace, it will return a rejected message. The remote node will include the origin node in its awareness of interaction record with a value equal to Peripheral. This awareness could be change to Full if the NimbusSpace is modified.

Having a look at the Figure 3 it is possible to appreciate the sequence of steps to carry out a load-balancing operation:

1. The node receives a message from the grid client containing the composition of processes to be executed.
2. Before starting the load-balancing operation, it will calculate the *Interactive Pool*.

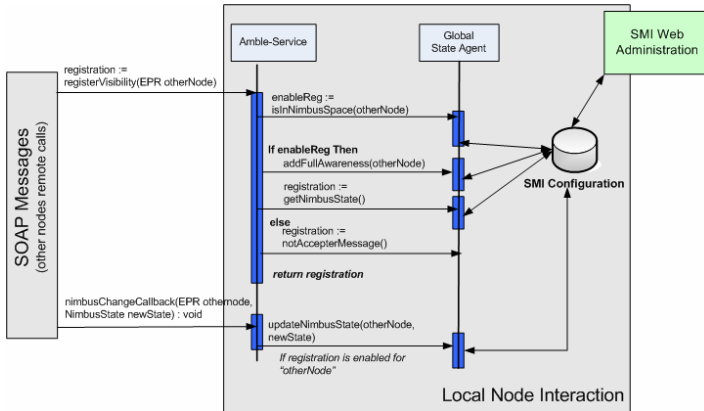


Fig. 2. General interaction diagram: AwareInt=Full and AwareInt=Peripheral

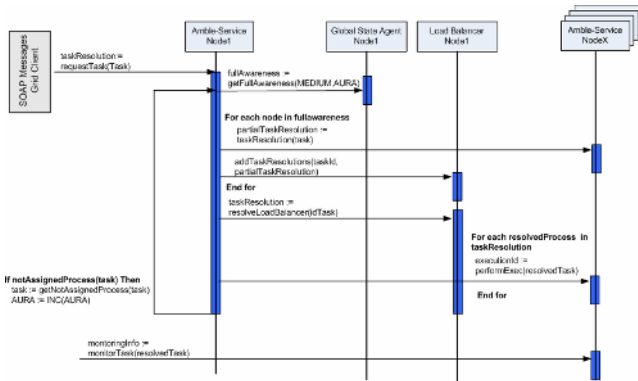


Fig. 3. General interaction diagram: Load-Balancing operation

3. The origin node asks to each of the nodes in the interactive pool (concurrently and through SOAP messages) for the score of each of the process of the T task.
4. The origin node will set this information in a list, and this list will be the input for the resolveLoadBalance function which implements the final selection heuristic and returns the assignment of the processes to the nodes.
5. For each of these assignments the origin node will send an execution message to the node. Even although the remote node could reject the execution of the process (its NimbusState would change to Null), having therefore the possibility of accepting the process execution. If later, the executionId assigned is NOT_ASSIGNED and the process will be assigned in the next round.
6. If any of the processes has not assignment, the SMI Engine will repeat the previous algorithm increasing the aura to catch those nodes that are located to a distance higher. This loop will be repeated until there are not more nodes in the system and, if later, those processes that have not been assigned will pass to the queue.

5 WS-AMBLE Architecture

WS-AMBLE provides an open interface to manage different levels of awareness, allowing different Virtual Organizations to share computational resources based on open protocols and interfaces. As far as we know, none of the last WS specifications offers functionalities useful enough as to create awareness models and none of the last WS specifications offers specific functionalities to manage task balancing delivery in collaborative grid environments, as WS-AMBLE. By means of WS-AMBLE, it will be possible to create new architectures oriented toward services. In Fig. 4 it is possible to appreciate the different levels of functionality offered by each of WS-AMBLE components:

1. **WS-Addressing:** It allows a transport mechanism to address services and messages (W3C standard)
2. **WS- Resource Framework:** This specification, evolution of the OGSF specification, has been recently standardised by OASIS consortium. It promotes an Standard way of deploying resources and how to consult about them.
3. **WS-Notification and WS-Topic:** Jointly, they provide the facility to establish mechanisms based on the model of interaction, publication and subscription.

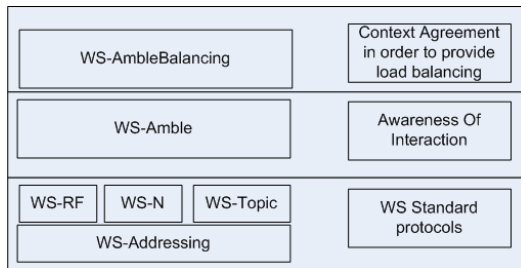


Fig. 4. WS-AMBLE Architecture

WS-AMBLE uses Standard mechanisms based on WS to interact with other resources. These mechanisms are supplied by the recently standardised WS-Resource Framework specification. On the other hand, the communication model, founded on publication/subscription, is based on the WS-N / WS-Topic specification. WS-N supplies the mechanism to subscribe two resource management nodes, in an AMBLE-grid environment, to the NimbusSpace as well as to the NimbusSpace.

6 AMBLE Experimental Results

The tests presented in this section have been evaluated in a real and heterogeneous services oriented grid environment. The system heterogeneity is reflected in the architecture of the computational nodes as well as in the OS. The grid infrastructure was made up for 20 nodes with the following characteristics: 8 of them were Intel Centrino P4 1.5 GHZ with 0.5 GB of memory ("Type A"), 11 of them were Intel P4 3.0 GHZ (B) with 1GB of memory ("Type B") and one Intel Xeon 3.6 GHz (C) with

4GB of memory (“Type C”). In order to carry out the model evaluation we have selected a set of CPU-intensive tests based in iterations over the Linpack benchmark [16]. Three different scenarios were used to make this evaluation. In each of these scenarios, there is a task T, composed of 20 processes, and a node N that receives the T execution request. Each of these scenarios also presents 4 different tests; each of them differs to the others in the size of the processes. The table 1 presents the response times, in seconds, for each of the tests executed in the different grid nodes.

Table 1. Response time

	Type A	Type B	Type C
Test 1	4,95	3,86	3,48
Test 2	24,02	19,20	17,80
Test 3	47,00	38,72	34,97
Test 4	232,93	192,30	175,03

Scenario A: The node N has full awareness of interaction with the rest of the nodes making up the grid, and therefore this node throws a load balancing operation to carry out the task execution taking into account all the nodes composing the grid. Table 2 presents the total response times of the system using the AMBLE implementation as well as the speedup of the AMBLE model related to each of the types of nodes.

Table 2. Global communication overhead and speedup related to the types nodes for scenario A

Amble	Communication overhead	Speed up vs A	Speed up vs B	Speed up vs C
5,9	5,65	0,84	0,65	0,59
6,964	5,76	3,45	2,76	2,56
8,233	5,88	5,71	4,70	4,25
17,69	5,92	13,17	10,87	9,89

Scenario B: This scenario raises the non ideal situation in which all the nodes in the grid are been infrautilised, but they are located in different auras. The node N has 10 more nodes inside its aura with a distance equal to 1 and the nine remaining in another aura with a distance equal to 2. The table 3 presents the speedup of the AMBLE model related to the local execution on each of the types of nodes as well as the communication overhead.

Table 3. Global communication overhead and speedup related to the types nodes for scenario B

Amble	Communication overhead	Speed up vs A	Speed up vs B	Speed up vs C
6,54	6,29	0,76	0,59	0,53
7,45	6,25	3,22	2,58	2,39
9,906	7,56	4,74	3,91	3,53
19,28	7,51	12,08	9,97	9,07

Scenario C: In the last scenario all the nodes in the grid are infrautilised but they are located in different auras. The grid client requests the execution of one task in the node N. This node has 10 more nodes inside its aura (aura1) and the nine remaining in another aura (aura2). However, in this situation half of the nodes that are inside the aura1 reject the execution of the processes assigned. Then, the load balancing algorithm increases the aura2 and therefore the other 9 remaining nodes could accept any of the processes that are looking for a location. The task delivery process is done among the nodes located in the aura2. While this distribution is been done, some of the nodes that are located inside the aura1 change their NimbusState and they could received new processes. The system will inform of this change and those processes that were not assigned among the nodes located in the aura2, will be assigned among all those nodes changed its NimbusState in the aura1. The table 4 shows the speedup and the communication overhead.

Table 4. Global communication overhead and speedup related to the types nodes for scenario C

Amble	Communication overhead	Speed up vs A	Speed up vs B	Speed up vs C
10,92	10,68	-7,44	0,45	0,35
12,08	10,88	5,72	1,99	1,59
14,13	11,78	20,84	3,33	2,74
24,29	12,52	150,74	9,59	7,92

The experimental results obtained are very successful and corroborate the usefulness of the AMBLE model as to be applied to work-load balancing operations in real heterogeneous grid environments. The performance improvements obtained by using this model are excellent in all the scenarios and in, almost, all the tests. It is worthy to highlight that the results achieved in the test1 in which the model get worse experimental results. These results are consequence to the small size of the task to be executed, which provokes a considerable communication overhead in this delivery operation, increasing the response times and making that these times were higher that the one associated to the local execution. Due to this fact, the speedup of these experiments are lower than 1. The results corresponding to the other tests show that the speedup experiment an important improvement. The communication overhead is the factor limiting the performance increase. This overhead is independent of the problem size. In this way, when the problem size increases, the parallelizable portion of the task also increases and therefore the speedup experiment a considerable improvement.

Finally, scenario A gets the best speedup results, related to the local execution. This is a consequence of the ideal conditions, for the execution of the AMBLE model, in which this scenario takes place. The scenario B presents an increment in the aura, and in scenario C there are some modifications on the NimbusState of some of the nodes. These situations imply that the load balancing model will require a set of messages to carry out the delivery operation and this communication overhead is reflected in the speedup results. However, in spite of this additional communication overhead the results are still very successful.

7 Conclusions

This paper presents an specification of an awareness model for balancing the load in collaborative grid environments in a collaborative multi-agent system. This model, called AMBLE (Awareness Model for Balancing the Load in Collaborative Grid Environments), extends and reinterprets some of the key concepts of the most successful models of awareness in Computer Supported Cooperative Work (CSCW), called the Spatial Model of Interaction (SMI). AMBLE manages the interaction in the environment allowing the autonomous, efficient and independent task allocation in the environment. The AMBLE implementation is based on the Web Services specifications and follows one of the key principles in the grid theory: the use of open, standard and public interfaces.

This model has been evaluated in a real and heterogeneous grid infrastructure. Different scenarios were designed for this purpose. The most important conclusions that could be extracted from the experimental results presented in this paper are: Firstly, the AMBLE model can contribute to the performance of heterogeneous systems by distributing the work-load in an equilibrating way among all the nodes composing the grid; Secondly, the communication overhead in a grid environment is a factor to be considered due to the remarkable limitations in the performance improvements. This overhead doesn't depend on the problem size, it mainly depends on the dynamism of the grid system, in each and every moment, and therefore it can not be predict beforehand. Finally, it is important to highlight that the size of the processes, to be distributed in the grid, has a fundamental impact in the global performance of the system. Those processes whose response time is low, are not suitable to be distributed in the grid because the communication overhead could be bigger that the local response time, entailing a worsening of the system

References

- [1] M. Beltrán, J. L. Bosque, A. Guzmán. *Resource Dissemination policies on Grids*. On the Move to Meaningful Internet Systems 2004: OTM 2004. Lectures Notes in Computer Science. Springer-Verlag. pp 135 – 144. October 25-29, 2004
- [2] Benford S.D., and Fahlén L.E. A Spatial Model of Interaction in Large Virtual Environments. Published in Proceedings of the Third European Conference on Computer Supported Cooperative Work (ECSCW'93). Milano. Italy. Kluwer Academic Publishers, pp. 109-124, 1993.
- [3] J. Cao et al., "Agent-Based Grid Load Balancing using Performance-Driven Task Scheduling", Proc. of the International Parallel and Distributed Processing Symposium 2003.
- [4] K. Chow and Y. Kwok, "On Load Balancing for Distributed Multiagent Computing", IEEE Transactions on Parallel and Distributed Systems 13(8), 787-801, Aug. 2002.
- [5] Fahlén, L. E. and Brown, C.G., *The Use of a 3D Aura Metaphor for Computer Based Conferencing and Teleworking*, Published in Proceedings of the 4th Multi-G Workshop, Stockholm-Kista, pp. 69-74, 1992.
- [6] Foster and C. Kesselman and J. Nick and S. Tuecke. *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. Globus Project, 2002.

- [7] Foster, N. R. Jennings and C. Kesselman, "Brain Meets Brawn: Why Grid and Agents Need Each Other", Proceedings 3rd Int. Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2004), New York, USA, 2004.
- [8] Galstyan, K. Czajkowski and K. Lerman, "Resource Allocation in the Grid using Reinforcement Learning", International Conference on Autonomous Agents and Multiagent Systems, 2004.
- [9] Greenhalgh, C., *Large Scale Collaborative Virtual Environments*, Doctoral Thesis. University of Nottingham. October 1997.
- [10] T. Kunz, "The influence of different workload descriptions on a heuristic load balancing scheme," IEEE Transactions on Software Engineering, vol. 17, no. 7, pp. 725–730, July 1991.
- [11] W. Shen et al., "Adaptive Negotiation for Agent-Based Grid Computing", Journal of the American Statistical Association, 2002.
- [12] Z. Zhang and S. Luo, "Constructing Grid System with Mobile Multiagent", Proc. of the Second Int. Conference on Machine Learning and Cybernetics, Xi'an, Nov. 2003.
- [13] <http://www.w3.org/>. Consulted in 2006.
- [14] <http://www.w3.org/2002/ws/>. Consulted in 2006.
- [15] <http://www.w3.org/TR/soap/>. Consulted in 2006
- [16] <http://www.netlib.org/benchmark/performance.pdf> Performance of Various Computers Using Standard Linear Equations Software. Technical Report CS-89-85, University of Tennessee, 2006.

An Agent System for Automated Web Service Composition and Invocation

In-Cheol Kim and Hoon Jin

Department of Computer Science, Kyonggi University
Suwon-si, Kyonggi-do, 442-760, South Korea
{kic, jinun}@kyonggi.ac.kr

Abstract. Semantic web services have the potential to change the way knowledge and business services are consumed and provided on the Web. The current semantic web service architectures, however, do not provide with integrated functionality of automated composition, dynamic binding, and invocation. Openness and dynamics of the Web environment limits the usage of previous approaches based upon the traditional AI planning techniques. This paper introduces a BDI agent system for semantic web service composition and invocation. Through some tests on healthcare web services, we found our agent-oriented approach has the potential enough to improve robustness and flexibility of semantic web services.

1 Introduction

Semantic web services augment web services with rich formal descriptions of their capabilities, thus facilitating automated discovery, composition, dynamic binding, and invocation of services within an open environment. Semantic web services, therefore, have the potential to change the way knowledge and business services are consumed and provided on the Web. Enabling semantic web services needs the infrastructure such as standard service ontology and architecture. The service ontology, such as OWL-S, aggregates all concept models related to the description of a semantic web service, and constitutes the knowledge-level model of the information describing and supporting the usage of the service. Currently several tools supporting OWL-S have been developed: OWL-S Editor, Matchmaker, Broker, Composer, and Virtual Machine [4]. However, the current semantic web service architectures do not provide with integrated functionality of composition, dynamic binding, and invocation [1]. Although there are some efforts for automated semantic web service composition, most of them are based upon the traditional AI planning techniques. Due to openness and dynamics of the Web environment, the web services composed through off-line planning are subject to fail. In order to overcome the drawback, this paper introduces a BDI agent system for semantic web service composition and invocation. Using some examples of healthcare web services, we test the feasibility of our approach.

2 BDI Agent Architecture

The most commonly used architecture for software agents is the *Belief-Desire-Intention* (BDI) model. Fig. 1 shows a PRS-based BDI architecture called JAM [3]. Each JAM agent is composed of five primary components: a *world model*, a *goal set*, a *plan library*, an *interpreter*, and an *intention structure*. The world model is a database that represents the beliefs of the agent. The goal set is a set of goals that the agent has to achieve. The plan library is a collection of plans that the agent can use to achieve its goals. The interpreter is the agent's "brain" that reasons about what the agent should do and when and how to do it. The intention structure is an internal model of the agent's current goals and keeps track of the commitment to, and progress on, accomplishment of those goals.

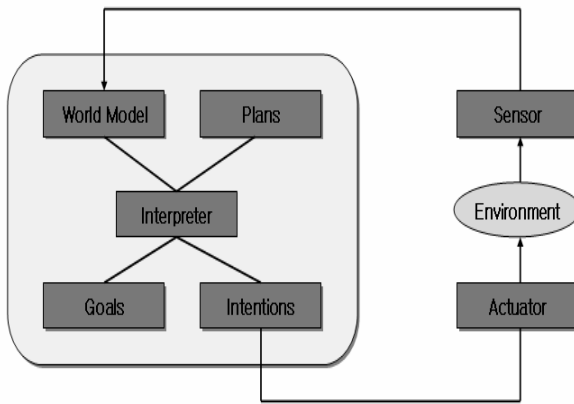


Fig. 1. JAM: A BDI Agent Architecture

The BDI architecture integrates traditional goal-directed reasoning and reactive behavior. Because most traditional deliberative planning systems formulate an entire course of action before starting execution of a plan, these systems are brittle to the extent that features of the world or consequences of actions might be uncertain. In contrast, the BDI architecture continuously tests its decisions against its changing knowledge about the world, and can redirect the choices of actions dynamically while remaining purposeful to the extent of the unexpected changes to the environment.

3 SWEEP II System

SWEEP II is a BDI agent system supporting automated semantic web service composition and invocation. As shown in Fig. 2, the core component of SWEEP II is the *JAM BDI engine*. Additionally, SWEEP II includes several main components such as *service description manager*, *OWL-S2JAM converter*, *ontology manager*, *reasoner*, *query processor*, *mediator*, *task manager*, *web service invoker*. The service description manager retrieves and stores OWL-S descriptions from the semantic web service repository.

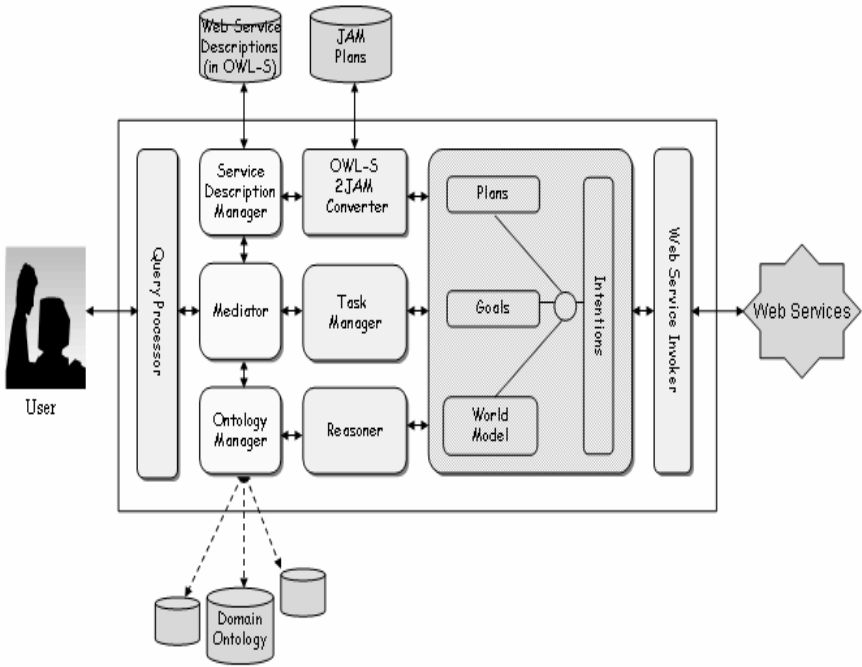


Fig. 2. Architecture of the SWEEP II System

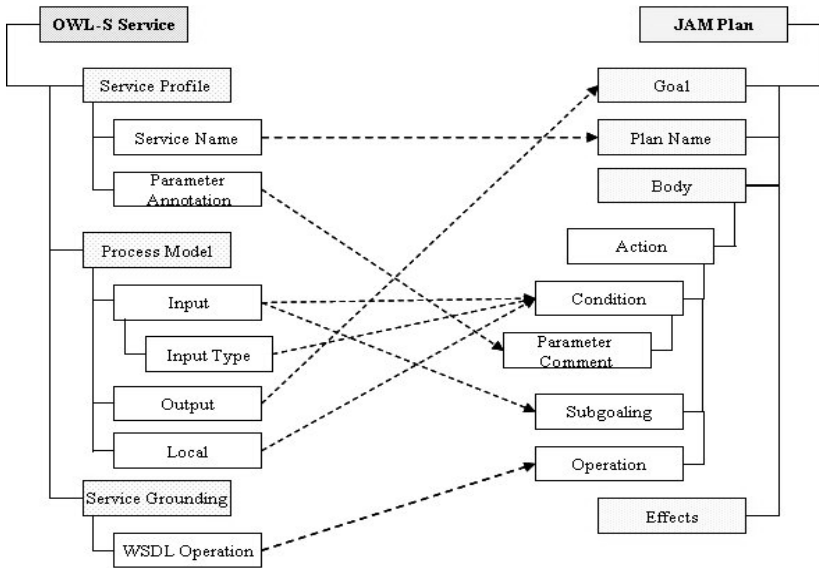


Fig. 3. OWL-S to JAM Plan Mapping

The OWL-S2JAM converter transforms the retrieved OWL-S descriptions into the corresponding JAM primitive plans. Based upon the domain ontologies managed by the ontology manager, the reasoner provides reasoning services for service matching, validation, and mediation. The query processor takes a request from the user. With the help of the mediator, it generates a task to be accomplished by JAM. In order to accomplish the given task, JAM reactively elaborates a complex plan from primitive plans considering its goals and the world model. According to the intentions of JAM, the web service invoker calls the atomic web services in order.

```

<?xml version="1.0"?>
  <owl:Ontology rdf:about="">
    ...
    <owl:imports rdf:resource="http://localhost/resources/ontology/Healthcare.owl"/>
  </owl:Ontology>
  <process:AtomicProcess rdf:ID="DiseaseSearchProcess">
    <process:hasOutput>
      <process:Output rdf:ID="disease">
        <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
          http://localhost/resources/ontology/Healthcare.owl#Disease
        </process:parameterType>
      </process:Output>
    </process:hasOutput>
    <process:hasInput>
      <process:Input rdf:ID="pain name">
        <process:parameterType rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
          http://localhost/resources/ontology/Healthcare.owl#Pain
        </process:parameterType>
        <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
          It's a name representing about pain.</rdfs:comment>
        </process:Input>
      </process:hasInput>
      <process:name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
        DiseaseSearchProcess</process:name>
      <service:describes>
        <service:Service rdf:ID="DiseaseSearch">
          <service:providedBy rdf:resource="#DiseaseSearch"/>
          ...
          <service:supports>
            <grounding:WsdGrounding rdf:ID="DiseaseSearchGrounding">
              <grounding:hasAtomicProcessGrounding>
                <grounding:WsdAtomicProcessGrounding rdf:ID="DiseaseSearchProcessGrounding">
                  <grounding:owlsProcess rdf:resource="#DiseaseSearchProcess"/>
                  <grounding:otherReference
                    rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
                    http://localhost:8080/axis/services/UFdiseaseSearch2?wsdl
                  </grounding:otherReference>
                  ...
                </grounding:WsdAtomicProcessGrounding>
              </grounding:hasAtomicProcessGrounding>
            </grounding:WsdGrounding>
          </service:supports>
        </service:Service>
      </service:describes>
    </process:AtomicProcess>
  </rdf:RDF>

```

Fig. 4. An Example of OWL-S Description


```

plan: {
NAME: "Disease Search"
GOAL: ACHIEVE disease_known "True";
PRECONDITION: RETRIEVE disease_known $known; (== $known "False");
BODY:
  ASSIGN $service_name "Disease Search";
  ASSIGN $currServ_paramNumIn 1;
  RETRIEVE simul_mode $smode;
  OR
  { TEST(== $smode "OFF");

  ASSIGN $cmt_param "Pain: ";
  ASSIGN $param (com.irs.jam.primitives.RequestUserInput.execute $cmt_param $param);
  WHEN: TEST (!= $param "") { ASSIGN $pain_name $param; };
  WHEN: TEST (== $param "") {
    DO{ OR{ ACHIEVE pain_name_known "True";
            RETRIEVE pain_name_known $known; }
        { EXECUTE printIn "Subgoaling Error!"; };
        } WHILE: TEST(!= $known "True" );
    RETRIEVE pain $pain_name; }; };
  ASSIGN $service_WSDL "http://localhost:8080/axis/services/DiseaseSearch?wsdl";
  EXECUTE com.irs.jam.primitives.ServiceCall.execute $service_name $service_WSDL
  $pain_name $disease;
  WHEN: TEST (!= $disease "") { ASSERT disease $disease; }; }
  { TEST(== $smode "ON");
  DO{ OR{ ACHIEVE pain_known "True";
          RETRIEVE pain_known $known; }
      { EXECUTE printIn "Subgoaling Error!"; };
      } WHILE: TEST(!= $known "True");
  SUCCEED; };
EFFECTS:
  EXECUTE com.irs.jam.primitives.GetCurrentPlanName.execute $service_name $planList;
  UPDATE (disease_known)(disease_known "True");
}

```

Fig. 5. The Corresponding JAM Plan

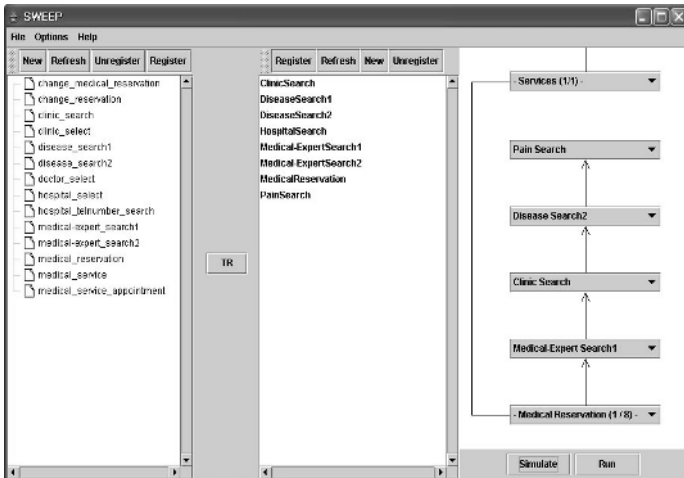


Fig. 6. A Screenshot of the SWEEP II System

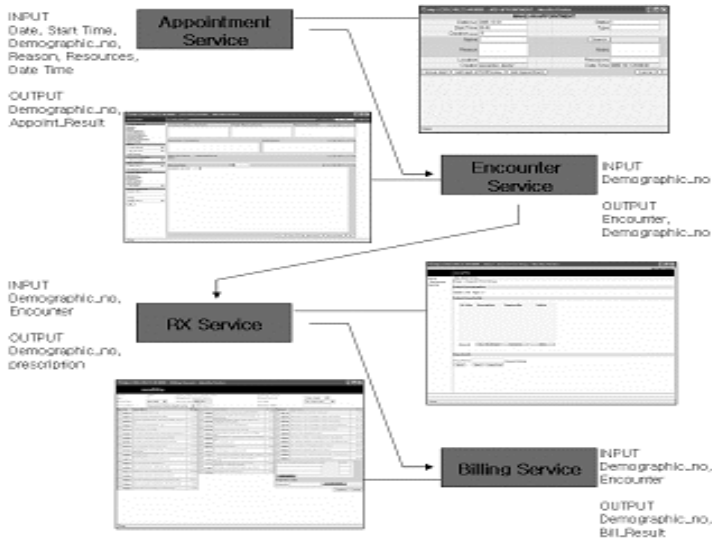


Fig. 7. Healthcare Web Services

Fig. 3 illustrates the mapping relationship from an OWL-S service description to the corresponding JAM plan. Outputs and effects of the OWL-S process are mapped to goals and effects of the JAM plan. Inputs and preconditions of the OWL-S process are mapped to subgoals and context of the JAM plan. The grounding WSDL operations of the OWL-S service are mapped to the external functions of the JAM plan. Many ontological concepts and properties in the OWL-S description are converted into the corresponding relations within the JAM world model. Fig. 4 shows an example of OWL-S service description and Fig. 5 represents the corresponding JAM plan.

Due to features of the embedded JAM engine, the SWEEP II system can adapt its decisions on web service composition and invocation against the changing Web environment. In this sense, the SWEEP II system realizes *dynamic composition* of semantic web services. Fig. 6 shows a screenshot of the SWEEP II system.

In order to test the feasibility of our SWEEP II system, we developed some examples of healthcare web services shown in Fig. 7. They are appointment service, encounter service, RX service, and billing service for patients. These are typically common services provided by many hospital information systems (HIS). For our purpose, however, we implemented them as OWL-S semantic web services. Through some tests on these semantic web services, we found that our SWEEP II system shows high robustness against unexpected failures and delays of web services.

4 Conclusions

We introduced SWEEP II, which is a BDI agent system for semantic web service composition and invocation. The core component of SWEEP II is the JAM BDI engine. The JAM BDI architecture continuously tests its decisions against its changing

knowledge about the world, and can redirect the choices of actions dynamically. Through some tests on healthcare web services, we found that our SWEEP II system can help to improve robustness and flexibility of semantic web services.

References

1. Cabral, L., Domingue, J., Motta, E., Payne, T., Hakimpour, F.: Approaches to Semantic Web Services: An Overview and Comparisons. Proceedings of the 1st European Semantic Web Symposium (ESWS2004), Springer-Verlag (2004) 225-239
2. Dickinson, I., Wooldridge, M.: Agents are not (just) Services: Investigating BDI Agents and Web Services. Proceedings of the Workshop on Service-Oriented Computing and Agent-Based Engineering (SOCABE'05) (2005)
3. Neto, R., Udipi, Y., Battle, S.: Agent-Based Mediation in Semantic Web Service Framework. Proceedings of the 1st AKT Workshop on Semantic Web Services (AKT-SWS04) (2004)
4. Marcus, J.: JAM: A BDI-theoretic Mobile Agent Architecture. Proceedings of the 3rd International Conference on Autonomous Agents (Agents'99) (1999) 236-243
5. McDermott, D.: Estimated-Regression Planning for Interactions with Web Services. Proceedings of the International Conference on AI Planning and Scheduling (AIPS'02), Morgan Kaufmann (2002)
6. Paolucci, M., Sycara K.: Autonomous Semantic Web Services. IEEE Internet Computing, Vol.7(5). (2003) 34-41
7. Paolucci, M., Soudry, J., Srinivasan, N., Sycara K.: A Broker for OWL-S Web Services. Proceedings of 2004 AAAI Spring Symposium on Semantic Web Services, MIT Press (2004)
8. Paolucci, M., Ankolekar, A., et al.: The DAML-S Virtual Machine. Proceeding of 2nd International Semantic Web Conference (ISWC2003), Springer-Verlag (2003) 290-305
9. Sirin, E., Parsia, B.: Planning for Semantic Web Services. Proceedings of 3rd International Semantic Web Conference (ISWC'04) Workshop on Semantic Web Services (2004)
10. Srivasta, B, Koehler, J.: Planning with Workflows – An Emerging paradigm for Web Service Composition. Proceedings of ICAPS 2004 Workshop on Planning and Scheduling for Web and Grid Services (2004)

An Auction-Based Semantic Service Discovery Model for E-Commerce Applications

Vedran Podobnik¹, Krunoslav Trzec², and Gordan Jezic¹

¹ University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Telecommunications, Unska 3, HR-10000 Zagreb, Croatia

² Ericsson Nikola Tesla, R&D Center, Krapinska 45, HR-10000 Zagreb, Croatia
vedran.podobnik@fer.hr, krunoslav.trzec@ericsson.com,
gordan.jezic@fer.hr

Abstract. Mediation between buyers (service requester's agents) and sellers (service provider's agents) is one of the most difficult problems in real electronic markets. In this paper, we propose an economic approach to solving this problem combined with AI (*Artificial Intelligence*) concepts. Firstly, we enable provider agents to dynamically and autonomously advertise semantic descriptions of available services by proposing a new auction model based on Pay-Per-Click advertising auctions. We call it the Semantic Pay-Per-Click Agent (SPPCA) auction. Requester agents then use two-level filtration of the advertised services to efficiently discover eligible services. In the first level of filtration, a semantic-based mechanism for matchmaking between services requested by buyers and those advertised by sellers is applied. Services which pass the first level of filtration are then considered on the second level. Here information regarding the actual performance of service providers is considered in conjunction with the prices bid by service provider's agents in the SPPCA auction. A final set of advertised services is then chosen and proposed to the buyer agent as an answer to its request.

Keywords: Intelligent software agents, Web services, the semantic Web, OWL-S, Pay-Per-Click (PPC) auctions, digital economy, electronic commerce.

1 Introduction

The initial architecture of the Web was geared towards delivering information to humans visually. At this very moment we are witnessing the transformation of the architecture of the Internet as it becomes increasingly based on goal directed applications which intelligibly and adaptively coordinate information exchanges and actions¹. Hence, the Internet is transforming into a medium which enables the so-called digital economy. The digital economy, by proliferation of the use of the Internet, provides a new level and form of connectivity among multiple heterogeneous ideas and actors, giving rise to a vast new range of business combinations [1]. Additionally, the digital economy automates business transactions by utilizing the technologies of Web services, the semantic Web and intelligent software agents.

¹ Source: IBM.

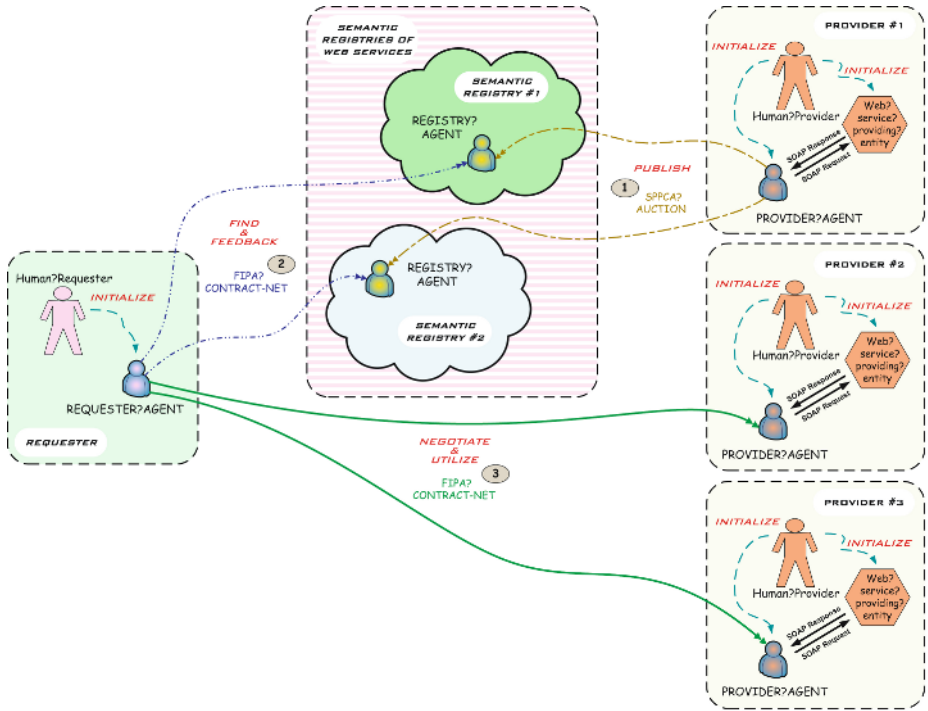


Fig. 1. The architecture of the modeled agent-mediated e-market of semantic Web services

The connection between AI (*Artificial Intelligence*) and economics has received a lot of attention recently [2]. The ideas presented in this paper are based on that connection. The automation of business processes is facilitated with the use of Web services and enabled by applying AI concepts realized through the technologies of intelligent software agents and the semantic Web. To the best of our knowledge, a model for efficient automated advertising of semantic Web service descriptions has not yet been developed. Thus, efficient business transactions are supported by modeling a new auction mechanism based on Pay-Per-Click advertising auctions, but adapted for agent environment and enhanced with a semantic dimension. We refer to this new auction model as the Semantic Pay-Per-Click Agent (SPPCA) auction.

We propose to improve the agent-mediated electronic market (e-market) of semantic Web services presented in [3] (shown in Figure 1) by upgrading its mechanism for matching service requesters with suitable service providers. Today’s semantic-enabled e-markets facilitate the discovery of eligible Web services solely through semantic matchmaking between descriptions of requested and advertised Web services [4]. Such an approach to service discovery can yield a large number of irrelevant search results since there is no assurance that information advertised by service providers is accurate [5]. Furthermore, services with identical descriptions may differ dramatically in performance levels [6].

In this paper we suggest automated semantic service discovery based, not only on service descriptions, but also on information about the actual performance of the

advertised Web services and the prices service provider's agents bid for advertising in the SPPCA auction. The performance of an advertised Web services is rated directly by the service requester's agent with respect to both price and quality. The proposed semantic service discovery model is not only interesting from scientific point of view, but is also very amenable to real e-commerce applications.

This paper is organized as follows. In Section 2 we give a brief summary of Web services, the Semantic Web and intelligent software agents. The proposed auction-based semantic service discovery model is presented in Section 3. The SPPCA auction is part of proposed service discovery model, so the auction's features are also described in Section 3. Section 4 proposes directions for future work and concludes the paper.

2 Technologies Enablers of the Digital Economy

2.1 Web Services

The integration of business applications is traditionally achieved by utilizing costly customized solutions for each business. Not only is it necessary to invest a great sum of money into the hardware infrastructure and software support needed for each new partner, but also a big effort is required to synchronize data formats and interaction protocols between partners [7]. As a result, the degree of reusability in current integration solutions is remarkably low. In this context, the Web provides an existing and highly available infrastructure for connecting business partners anywhere and anytime. In addition, Web services [8] support interoperable machine-to-machine interaction over a network by providing a set of standards for the provision of functionality over the Web. The problem is that descriptions of Web services (WSDL is the industry standard for Web service description) are purely syntactic, so the classical architecture of Web services [3] is hardly amenable to automation [7].

2.2 Semantic Web

Tim Berners-Lee's vision of the semantic Web [9] provides a foundation for the semantic architecture of Web services [10]. By applying the OWL-S (*Web Ontology Language for Services*), every Web service can be described with an ontology. Each OWL-S ontology utilizes one or more domain ontologies which define the concepts important for a particular domain of interest. Concepts in domain ontologies, as well as the relations between the concepts themselves, are specified with OWL (*Web Ontology Language*), a semantic markup language for publishing and sharing ontologies on the World Wide Web [3]. Therefore, the semantic architecture of Web services supports knowledge exchange among collaborating e-business actors in the digital economy. Ontologies provide a shared vocabulary to represent the meaning of entities while knowledge representation provides structured collections of information and inference rules for automated reasoning. As a result, intelligent software agents can interpret and exchange semantically enriched knowledge for users [11].

2.3 Intelligent Software Agents

An intelligent software agent is a program which acts on behalf of its owner while conducting complex information and communication actions over the Web. From the

owner’s point of view, agents improve their efficiency by reducing the time required to execute personal and/or business tasks. The goal of the technology of software agents is also to reduce the owner’s efforts while transacting business [12, 13].

3 The Semantic Service Discovery Model

The semantic service discovery model presented in this paper combines an economic approach to service discovery with AI concepts. Two-level filtration of advertised services is used by requester agents for efficient discovery of eligible services. First-level filtration is based on semantic matchmaking between descriptions of services requested by buyers and those advertised by sellers. Services which pass the first level of filtration are then considered in the second filtration step. Second-level filtration combines information regarding the actual performance of service providers and the prices bid by service provider’s agents. The performance of service providers (with respect to both price and quality) is calculated from requester agent’s feedback ratings. Following filtration, a final set of advertised services is chosen. This set is then recommended to buyers as an answer to their request.

Service providers’ agents put up bids for advertising their services in the Semantic Pay-Per-Click Agent (SPPCA) auction. The SPPCA auction is a new auction model we developed to enable provider agents to dynamically and autonomously advertise semantic descriptions of available services. The SPPCA auction is based on Pay-Per-Click advertising auctions which are currently used by Google and Overture (a search engine which provides sponsored search results for Yahoo, MSN, Lycos, AltaVista etc.) as a new mechanism for placing and paying for advertisements.

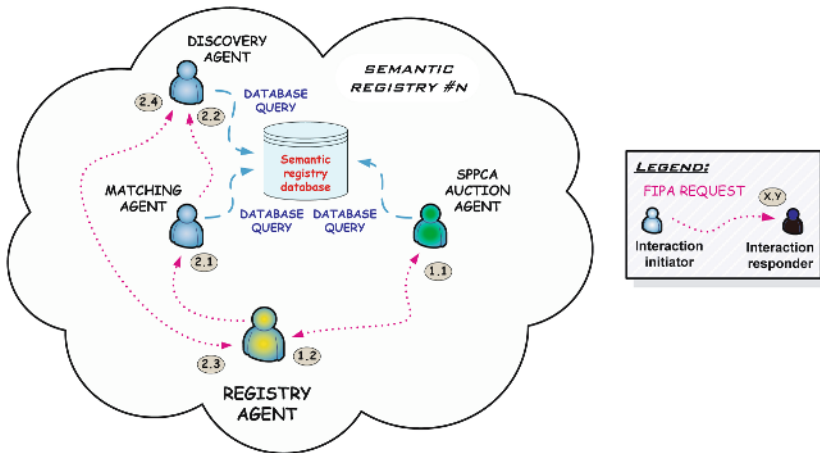


Fig. 2. The architecture of a semantic registry

Figure 2 shows the architecture of a semantic registry in the agent-mediated e-market of semantic Web services depicted in Figure 1. We can see that the Registry Agent serves as an interface agent between provider/requester agents and the semantic

registry. The SPPCA Auction Agent, the Matching Agent and the Discovery Agent enable necessary functionalities of the semantic registry. These agents are allowed to make queries to the registry's database. The SPPCA Auction Agent is in charge of conducting the SPPCA auction. Interaction 1.1 is used for registering/deregistering service provider agents in the auction, while the SPPCA Auction Agent uses interaction 1.2 to announce a new auction round. The Matching Agent facilitates semantic matchmaking which corresponds to the first level of filtration in the service discovery process. It receives OWL-S descriptions of requested services through interaction 2.1 and forwards a list of semantically eligible services through interaction 2.2 to the Discovery Agent which carries out second-level filtration and recommends top-ranked advertised services (interaction 2.3). Sometime later, the Discovery Agent receives feedback information from requester agents regarding the performance of the proposed services (interaction 2.4).

Figure 3 shows the structure of records of advertised service descriptions in the semantic registry database. Field 1 (a reference to an OWL-S description) is used to identify the advertised service and to retrieve its semantic service description if needed. Fields 2 (the quality rating of the service) and 3 (the price rating of the service) are used to create a performance model of the advertised service. Fields 4 (service's bid value in the current round of the SPPCA auction) and 5 (service's budget until the next round of the SPPCA auction) are utilized in the SPPCA auction.

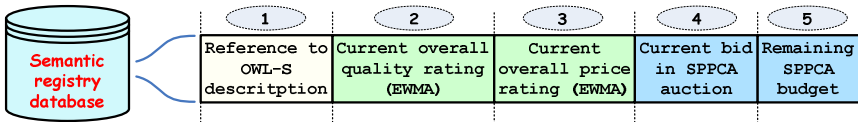


Fig. 3. The record of an advertised service description in semantic registry database

3.1 Semantic Matchmaking of Service Descriptions (The Matching Agent)

Most approaches suggested for semantic matchmaking use standard description-logic reasoning to automatically determine whether one service description matches another. Prominent examples of such logic-only based approaches to semantic service discovery are provided by OWL-S description matchmakers such as the Matching algorithm used in [3] (which is based on [14]), the OWLS-UDDI matchmaker [15], RACER [16], MAMA [17] and by the WSMO service discovery approach [18]. These approaches do not exploit semantics that are implicit, i.e. patterns and/or relative frequencies of service descriptions computed by techniques such as data mining, linguistics, or content-based information retrieval (IR). In order to exploit these techniques, our Matching Agent uses OWLS-MX [19], a hybrid semantic Web service matching tool which combines logic-based reasoning with approximate matching based on syntactic IR similarity computations. The objective of hybrid semantic Web service matching is to improve the performance of semantic service retrieval by appropriately exploiting both crisp logic-based and approximate semantic matching where each of them alone would fail.

3.2 The Performance Model of Service Providers (The Discovery Agent)

A performance model tracks a service provider's past performance which can be used to estimate its performance with respect to future requests [6]. Our model monitors two aspects of the advertised service's performance – the service quality and the price of utilizing the service. There is a difference in the way requester agents provide the semantic registry with feedback information regarding the quality and the price of particular advertised services. The requester agent gives the Registry Agent feedback regarding particular service's price (called the 'price rating') immediately after the discovery process ends. More precisely, this occurs after the requester agent has negotiated the conditions of utilizing a service with all recommended provider agents. Since the negotiation is automated, it only lasts for a few moments. Supplying the Registry Agent with feedback regarding the quality (called the 'quality rating') of particular service does not happen immediately after the service discovery process, but sometime in the future after the requester agent has used the chosen service. The rating value (both for quality and price) is a real number $r \in [0, 1.0]$. A rating of 0.0 is the worst (i.e. the service provider could not perform the service at all and/or utilizing the service is very expensive) while a rating of 1.0 is the best (i.e. the service provider performs the service perfectly and/or utilizing the service is very cheap).

The overall rating (fields marked as 2 and 3 in Figure 3) can be calculated a number of ways, but here we chose the EWMA (*Exponentially Weighted Moving Average*) method. The advantage of using EMWA is its adaptive nature, i.e. it can capture the trend of dynamic changes. Furthermore, it is computationally very simple since the new overall rating can be calculated from the previous overall rating and the current feedback rating (i.e. there is no need to store old ratings which is desirable due to scalability issues). EWMA is defined as follows:

$$\tilde{x}_t = \alpha x_t + (1 - \alpha)\tilde{x}_{t-1} \text{ for } t=1,2,\dots \quad (1)$$

Where:

- \tilde{x}_t is the new forecast value of x ;
- x_t is the current observation value (in our case, the new feedback rating) ;
- \tilde{x}_{t-1} is the previous forecast value ;
- $0 \leq \alpha \leq 1$ is a constant that determines the depth of memory of the EWMA. As the value of α increases, more weight is given to the most recent values.

3.3 Service Advertising in Pay-Per-Click Auctions (The SPPCA Auction Agent)

A Pay-Per-Click (PPC) advertising auction is an auction for sponsored positions in search engines. For instance, if a user types in a search for "Croatian hotels" on Google, he/she will get back a set of listings. These include sponsored sites which have paid on a PPC auction to have their companies shown.

PPC auctions run every minute of the day for every possible character sequence. In each auction, a competitor c enters a bid $b_k(c)$ which is the amount he/she is willing to pay should a customer click on his/her advertisement in the search results for keyword k . The auctioneer (e.g. Google) sorts the bids for keyword/auction k and awards position 1 to the highest bidder, position 2 to the second highest bidder, and so

on. The participant will then pay an amount equal to the number of customers that visit their web-site multiplied by their bid price [20].

Keyword auctions are an indispensable part of the business model of modern web search engines and are responsible for a significant share of its revenue [21].

Semantic Pay-Per-Click Agent Auctions. We have adapted PPC for the agent environment and enhanced it with a semantic dimension. We call this new auction model the Semantic Pay-Per-Click Agent (SPPCA) auction.

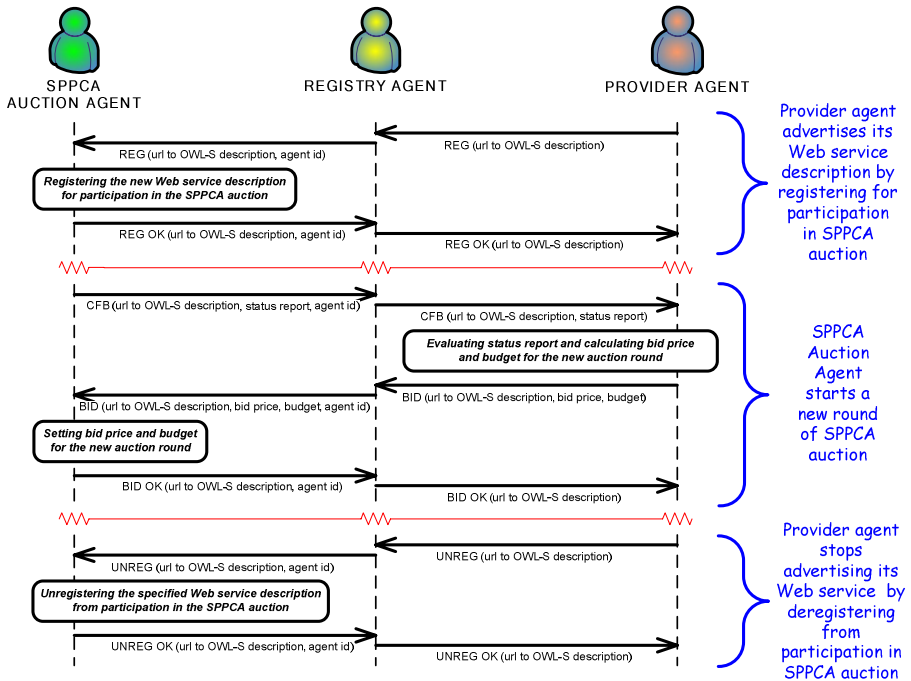


Fig. 4. SPPCA auction interactions

Figure 4 shows three different types of SPPCA auction interactions. The first type of interaction uses REG (*Registration*) and REG OK (*Registration successful*) messages to register a provider agent’s specific Web service for participation in the SPPCA auction. The second type of interaction uses CFB (*Call for Bid*), BID (*Bid*) and BID OK (*Bid successful*) messages to place bids in a new auction round. One of the parameters sent in CFB message is a status report sent at the end of each auction round. In such a report, the SPPCA Auction Agent sends to the service provider agent information regarding events related to its advertisement which occurred during the previous auction round. The most important information is that regarding how much of the provider agent’s budget was spent (i.e. the provider agent’s bid price multiplied with the number of recommendations of its service to requester agents). The provider agent also receives information about the final rankings of its advertisement in discovery processes in which the respective advertisement passed first-level filtering (i.e. semantic matchmaking). On the basis of this information, the service provider

agent can conclude whether its bids/budget was too high or too low compared to its competitors. It then sends a new bid (field 4 in Figure 3) and a new maximum budget value (field 5 in Figure 3) as part of its BID message. These values become relevant in the next auction round. If the provider agent's budget is spent before the end of the round, its advertisement becomes inactive until the end of that round and is therefore not considered in any of the matching processes. The third type of interaction uses UNREG (*Deregistration*) and UNREG OK (*Deregistration successful*) messages to deregister a provider agent's specific Web service from participation in the SPPCA auction.

A classic PPC auction runs for each particular character sequence and, thus, for every possible character sequence there is a separate auction. This model has a several shortcomings. First of all, there is a scalability problem. Namely, there are a huge number of syntactically valid combinations which result in a vast number of concurrent PPC auctions. Another problem is that separate auctions are held for synonyms (syntactically different words with the same meaning; e.g. car and automobile). From the service provider's point of view, it can be very complex and expensive for them to bid in auctions for all synonyms. From the service requester's point of view, it is very complicated to search all synonymous words when they require a particular service. The last disadvantage of the classic PPC auction model we consider here is competitor click fraud. This occurs when one company clicks on a competitor's advertisement to spend their budget with the long term aim of making PPC advertising too expensive for them and therefore removing them as a competitor from the search engine's results.

The auction model proposed in this paper, SPPCA, solves the shortcomings described above. The first problem of a vast number of concurrent auctions is solved by having one semantic registry running only one SPPCA auction and connecting provider agent's bids with their OWL-S descriptions and not a specific keyword. The second problem of running separate auctions for synonyms is solved by introducing the semantic Web technology which uses OWL-S descriptions for description of advertised services. The third problem of competitor click fraud cannot occur in the SPPCA auction model since a requester cannot predict which advertised services will be recommended as response to a request. Namely, the answer to each new discovery request is calculated dynamically and depends on fast-changing variables which are unknown to all entities outside the semantic registry. Hence, a requester agent cannot purposely cause the semantic registry to charge the targeted service provider agent by making a discovery request without the intent of utilizing any Web service.

3.4 Calculating Proposed Services

The proposed service discovery process (Figure 5) in its first step uses solely semantic matchmaking. In the second step it combines the SPPCA bid price and the current service performance rating to choose the final set of top-ranked advertised services which are then recommended to the requester agent. Since our performance model monitors two aspects of the advertised service's performance (i.e. its quality and price), the requester agent defines a weight factor which determines the significance of each of the two aspects in the process of calculating the final proposal.

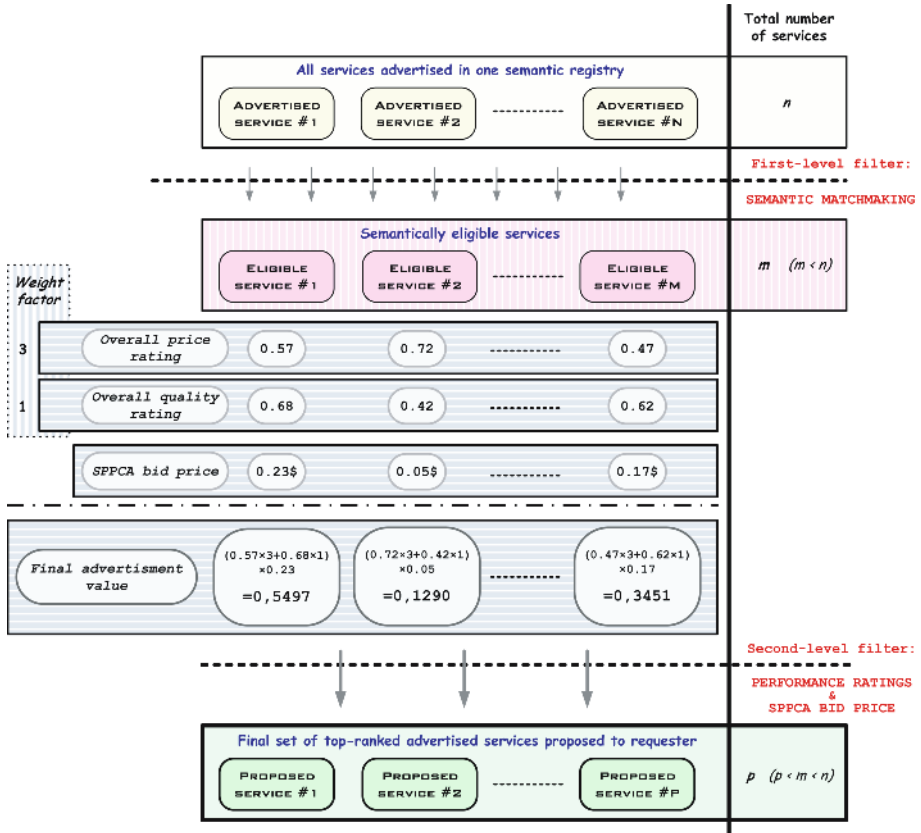


Fig. 5. The service discovery process

4 Conclusions and Future Work

In this paper, we propose a new model for auction-based semantic service discovery. The ideas presented combine economic and AI approaches. As a result, the suggested model is interesting, not only from a scientific point of view, but is also very amenable for real e-commerce applications. Two-level filtration of advertised services is used for the efficient discovery of eligible services. First-level filtration is achieved through semantic matchmaking. Second-level filtration is then applied by combining information regarding the actual performance of the service providers together with the prices bid by the service provider’s agents in the SPPCA auction. Following filtration, final sets of advertised services are then proposed to buyers as answers to their requests. SPPCA is a new auction model we have developed to enable provider agents to dynamically and autonomously advertise semantic descriptions of available services.

For future work we plan to test the proposed model from the aspects of scalability and time performance, as well as analyze its economic efficiency.

References

1. Carlson, B.: The Digital Economy – What is New and What is Not?. Structural Change and Economic Dynamics, Vol. 15, Elsevier (2004). 245-264
2. Wurman, P.R., Wellman, M.P., Walsch, W.E.: Specifying Rules for Electronic Auctions. AI Magazine, Vol. 23 (3), American Association for Artificial Intelligence (2002). 15-24
3. Podobnik, V., Jezic, G., Trzec, K.: An Agent-mediated Electronic Market of Semantic Web Services. In Proc. of the AAMAS Workshop on Business Agents and the Semantic Web (BASEWEB), Hakodate, Japan, 2006. 1-10
4. Srinivasan, N., Paolucci, M., Sycara, K.: An Efficient Algorithm for OWL-S Based Semantic Search in UDDI. In Proc. of the 1st Int. Workshop on Semantic Web Services and Web Process Composition (SWSWPC), San Diego, CA, USA, 2004. 96-110
5. Lim, W.S., Tang, C.S.: An Auction Model Arising from an Internet Search Service Provider. European Journal of Operational Research, Vol. 172, Elsevier (2006). 956-970
6. Luan, X.: Adaptive Middle Agent for Service Matching in the Semantic Web – A Quantitative Approach. Thesis, University of Maryland Baltimore County, 2004
7. de Bruijn, J., Fensel, D., Keller, U., Lara, R.: Using the Web Service Modeling Ontology to Enable Semantic E-business. Communications of the ACM, Vol. 48 (12), 2005. 43-47
8. W3C Web service architecture: <http://www.w3.org/TR/ws-arch/>
9. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American, Vol. 284 (5), 2001. 34-43
10. Singh, R., Iyer, L.S., Salam, A.F.: The Semantic E-business Vision. Communications of the ACM, Vol. 48 (12). 2005. 38-41
11. Hendler, J.: Agents and the Semantic Web. IEEE Intelligent Systems, Vol. 16 (2), 2001. 30-37
12. Podobnik, V.: Software Agents for Electronic Market. Thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, 2006 (in Croatian)
13. Bradshaw, J.M.: Software Agents. MIT Press, Cambridge, Massachusetts, USA (1997)
14. Tang, S.: Matching of Web Services Specifications using DAML-S Descriptions. Thesis, Technical University of Berlin, 2004
15. Sycara, K., Paolucci, M., Anolekar, A., Srinivasan, N.: Automated Discovery, Interaction and Composition of Semantic Web Services. Journal of Web Semantics, Vol. 1 (1), Elsevier (2004).
16. Li, L., Horrock, I.: A Software Framework for Matchmaking Based on Semantic Web Technology. In Proc. of the 12th Int. World Wide Web Conference (WWW), Budapest, Hungary, 2003. 331-339
17. Colucci, S., Noia, T.D., Sciascio, E.D., Donini, F., Mongiello, M.: Concept Abduction and Contraction for Semantic-based Discovery of Matches and Negotiation Spaces in an E-marketplace. Electronic Commerce Research and Applications, Vol. 4 (3), Elsevier (2005). 345-361
18. Keller, U., Lara, R., Lausen, H., Polleres, A., Fensel, D.: Automatic Location of Services. In Proc. of the 2nd European Semantic Web Conference (ESWC), Crete, Greece, 2005. Lecture Notes in Computer Science, Vol. 3532, Springer (2005). 1-16
19. Klusch, M., Fries, B., Sycara, K.: Automated Semantic Web Service Discovery with OWLS-MX. In Proc. of the 5th Int. Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Hakodate, Japan, 2006. 915-922
20. Kitts, B., LeBlanc, B.: Optimal Bidding on Keyword Auctions. Electronic Markets, Vol. 14 (3), 2004. 186-201
21. Aggarwal, G., Goel, A., Motwani, R.: Truthful Auctions for Pricing Search Keywords. In Proc. of the 7th ACM Conference on Electronic Commerce (EC), Ann Arbor, Michigan, USA, 2006.

Implementation of Business Process Requiring User Interaction

Guillermo López¹, Valeria de Castro², and Esperanza Marcos²

¹ Aldebarán Tecnologías de la Información. Chile, 4
28290 Las Rozas, Spain

guillermo.lopez@grupoaldebaran.com

² Universidad Rey Juan Carlos. Tulipán, s/n
28933 Móstoles, Spain

{valeria.decastro, esperanza.marcos}@urjc.es

Abstract. Innovations in web technologies during last years have caused changes in the way that Web Information Systems are developed. New tools such as XML or Web Services and new paradigms, such as Service Oriented Computing make IT analysts focus on business process. To implement these business process there are several technologies. The aim of this work is to analyze business process implementation possibilities by using existing technologies, but focusing on those that require a user interaction, such as: a process to buy a flight ticket or a process for billing approval. In this work, we follow a comprehensive method for business process development to get detailed business process models and then we analyze the implementation possibilities of the resulting models, describing advantages and disadvantages for each technology analyzed and pointing out a possible workaround.

Keywords: Service Oriented Computing, Model Driven Architecture, eGovernment, Web Information Systems, BPEL, BPEL4People.

1 Introduction

The increase of expectations in Web technology environments has lead to the need of implementing new more complex business processes, similar to the traditional application functionalities [1]. Web Information System Users (WIS Users) are not already content just showing information in static or dynamic HTML pages, but they demand the system a greater user interaction and the possibility of doing operations that are more complex. Powerful languages such as XLANG [2], Web Services Flow Language (WSFL) [3] and Business Process Execution Language for Web Services (BPEL4WS or BPEL in short) [4] have appeared to support these new requirements and to help in the design and implementation of business process. Nevertheless, business analysts have some problems using them, specially, when business processes include human-performed activities.

In this work, we use a comprehensive method for service composition modeling, which is focused on the behavior aspect of a WIS, to get a detailed business process model that will be used to analyze business process implementation possibilities. As a

case study, we present Atento. Atento is a real WIS used in a Spanish Public Administration to manage citizens' relationships solving the problem to access information anytime or anywhere and modifying citizens' service requests information stored in an external system. Examples of these relationships are the request and the issue of an official certificate or the inquiry for changing personal data.

The rest of the article is organized as follows: an overview of the model driven architecture of MIDAS is presented in Section 2. Section 3 describes the case study: Atento system, identifying the services offered to users. Section 4 illustrates the method and the analysis of implementation possibilities is explained in Section 5. Finally, Section 6 concludes underlying the main contribution and the future works.

2 MIDAS Methodology

This work forms part of the methodological framework of MIDAS, a methodological framework for the agile development of WIS [5] based on Model Driven Architecture (MDA) [6] proposed by the Object Management Group [7].

In the frame of MIDAS, it has been proposed a method for service composition modeling that is focused on the behavior aspect of a WIS [5].

MIDAS proposes to model the WIS according to two orthogonal dimensions (see Fig. 1): on the one hand, taking into account the platform dependence degree (based on the MDA approach) and specifying the whole system by Computation Independent Models (CIMs), Platform Independent Models (PIMs) and Platform Specific Models (PSMs). On the other hand, according to three basic aspects [8]: *hypertext, content and behavior*.

This work is focused on the behavioral aspect although focused on PIM level. As it is shown in Fig. 1, we use four different models (shaded in the figure): the *user services model*, the *extended use cases model*, the *service composition model* and the *service process model*. Our method introduces a new set of concepts, new models and mapping between them. All the steps needed to build these behavior models were described in previous works [5]. In the following section, we will illustrate these models through Atento system.

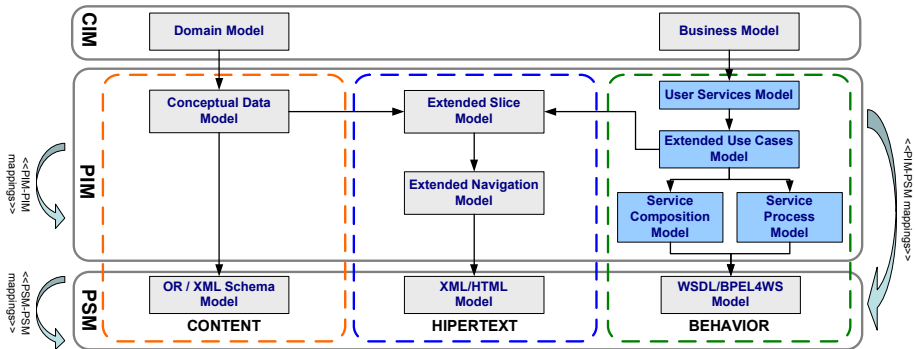


Fig. 1. Model Driven Architecture of MIDAS

3 A Case Study: Public Administration Automatization

Atento is a citizens relationships system that works over the intranet of a Spanish Public Administration. It allows officers accessing to the service request information sent by citizens to the Public Administration and the answers given to them. It does not matter which the source channel was (telephone, email, postal or public attention office) every service request is stored in the same place offering a 360-degree vision of the citizens and improving public attention level given to them. An example of citizen-public administration relationship is an information request submitted to know the steps needed to carry out a specific procedure. This information about citizens is not stored in Atento but in an external system “Citizen Relationship Management System”. This system is used by *phone* attention officers while Atento is used by *public* attention officers. Citizen Relationship Management System offers same functionality that Atento, but offers also more functionality necessary in phone attention processes.

Public Attention Officers with a valid username and password will be able to: a) query and modify citizens’ contact data, b) query and modify citizens’ service request, c) create new citizens’ service requests or d) save citizens’ satisfaction surveys. Fig. 2 shows Atento Architecture and relations between other external systems such as Citizen Relationship Management System.

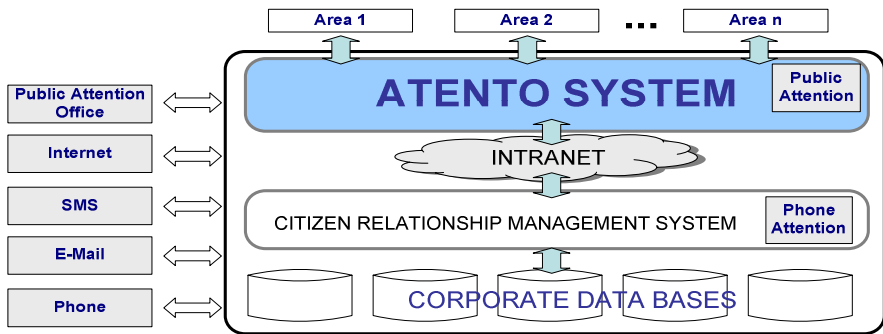


Fig. 2. Atento architecture

4 Atento System Modeling

4.1 Generating the User Services Model

The aim of this activity is building the *user services model*. This model shows the *services that the user really needs from the system*. These are known as *conceptual user services*.

All the modeling process is focused on these conceptual user services. To do a correct navigational model the focus must be on *the services that the user really requires from the system*. In *Atento*, in order to *Save a Citizen Service Request*, Public Attention Officer must first “login into system” by typing his username and password.

Then he will “search citizens” clicking on the proper button to find citizen to save the service request. He will finally “create a service request” clicking the proper button and typing the service request information. In our point of view, *Save a Citizen Service Request* is the service that the user really needs from the system in order to save a service request, that is, the conceptual user service. However, “login into system” is a required action to save a service request. The system has to guide user through mandatory actions in order to achieve his objectives.

The generation of this model begins *identifying users of the system (as actors)*. As we mentioned above, Atento is a WIS used by officers who need to access information stored on an external system, from now on, CRM System. Therefore, we identify one actor: Public Attention Officers. Public Attention Officer is the user who will interact with Atento system.

Secondly, we have to identify *the conceptual user services* taking into account the services required by the user. The real aims of a Public Attention Officer are: a) “*To Edit Citizen Data*”, b) “*To Save Citizen Service Request*”, c) “*To Query Service Request*” and d) “*To Save Citizen Satisfaction Survey*”. This model is represented using UML use case diagrams and identifying conceptual user services using <<CUS>> stereotype (see Fig. 3). This point onwards, we will focus only on the conceptual user service “*To Save Service Request*” to explain how to build the rest of the models we propose.

4.2 Generating the Extended Use Cases Model

The aim of this activity is building the *extended use cases model*. This model shows all the different activities and tasks that compose a conceptual user service from the previous model. These tasks are called *use services*. So, conceptual user service “*To Save Service Request*” splits into the following use services: Login, Search Citizens, Show Advanced Search Fields for Citizens, Show Citizens Search Results, Type Service Request Data and Show Recording Results.

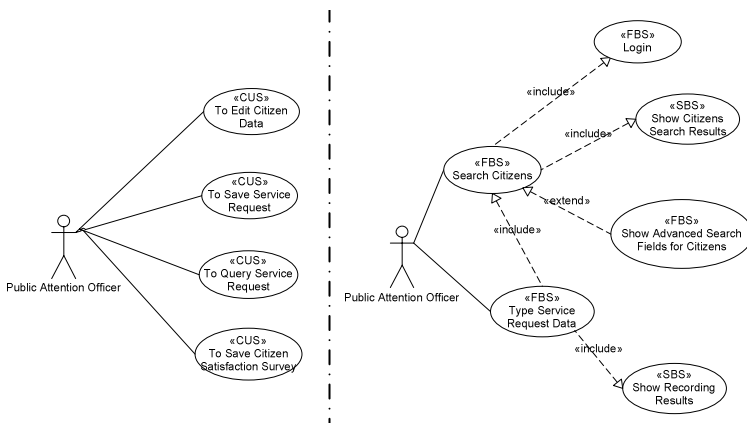


Fig. 3. User Service Model and Extended use case model for "To Save Service Request" service

After this, *each use service has to be identified as a composite* (use services composed by other basic or composite use services) *or a basic use service*. Secondly as *functional* (representing a basic functionality required by the WIS to carry out a service that implies some kind of interaction with the user, generally typing input data) or *structural* (those use services that provide a data view). We represent use services as use cases stereotyped by <<CS>> for composite use services, <<FBS>> for basic functional use services and <<SBS>> for basic structural use services.

Finally, we have to *identify include and extend relationships between use services*. In our example *Type Service Request Data* is a functional basic use service which is associated with *Show Recording Results* by an include relationship. Fig. 3 shows extended use case model for the *To Save Service Request* service.

4.3 Generating the Services Process Model

The aim of this activity is building the *services process model*. This model shows the execution flow of *service activities* needed to carry out the service.

Mapping each structural and functional basic service into service activities is necessary to build these models. Next step is *drawing the control flow between service activities* describing the service activities execution sequence.

Fig. 4 shows the service process model obtained for the *To Save Service Request* service.

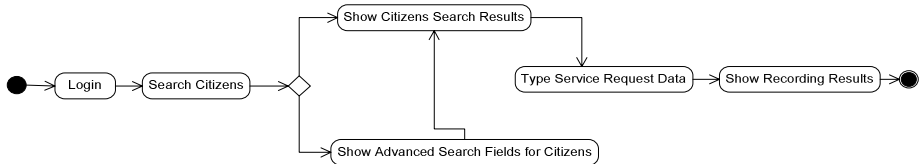


Fig. 4. Service process model for the "To Save Service Request" service

4.4 Generating the Services Composition Model

The aim of this activity is building the *services composition model*. Similar to previous one, this model represents *service process*, but in a detailed way that is, including two new concepts, *activity operation* and *business collaborator*. Activity operations are those entities that carry out some activities involved in the conceptual user service. Business collaborators are those entities internal or external to the system that carry out the activity operations. To build the model we will use as inputs the elements identified in the service process model and some of the elements identified in business model, such as system users and business collaborators.

First step is *identifying the business collaborators*. In our case study, we have detected two business collaborators, Atento system, which implements fundamental operations of the system; and the CRM system, storing information about citizens-public administration relationships. We use an UML activity diagram to draw the service composition model, *representing business collaborators as partitions* and external business collaborators stereotyped as <<external>>.

Second step is *splitting service activities, identifying activity operations supported by each service activity*. In our example: service activity *Show Citizens Search Results*

is divided into two activity operations: *Search Citizen* and *Show Citizens Search Result*. In the same way, the service activity *Type Service Request Data* is split into two activity operations: *Type Service Request Data* and *Save Service Request*. Finally, the service activity *Show Recording Results* is divided into two activity operations: *Show Save Successful Message* and *Show Save Error Message*. The rest of the service activities do involve only one activity operation. Once activity operations are identified, they must be distributed into partitions according with to the business collaborator that carries out the operation.

Third step is *identifying the activity operations that can be implemented as web services*. In the service composition model for *To Save Service Request* service, we have identified two activity operations that can be implemented as web services: *Search Citizen* and *Save Service Request*. Web services are stereotyped as <<WS>>.

Finally, we have to *identify control and data flow between activity operations*, taking into account control and data flow between services activities defined the in previous model. Control and data flow will be the same in both models, but if a service activity has been divided into several operations activities, control flow must be set by the designer.

Fig. 5 shows the results of this process for the “*To save Service Request*” service.

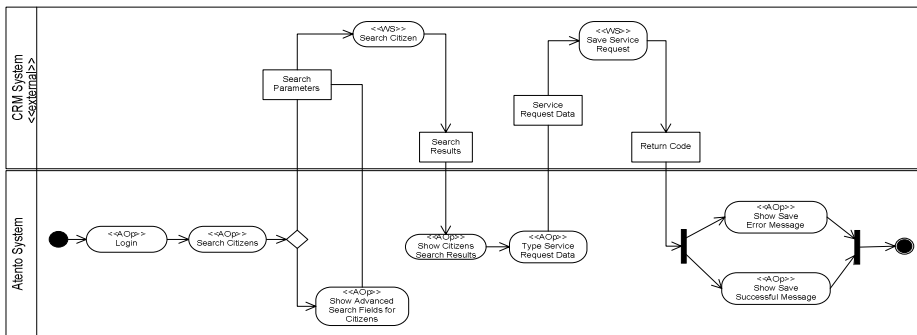


Fig. 5. Service composition model for "To Save Service Request" service

5 Analysis of Business Process Implementations Possibilities

We have analyzed implementation possibilities for the service composition model, transforming it to a specific Web technology. We have considered three different possibilities of implementation: BPEL, WS-Coordination/WS-Transaction or implement inside web application logic. These three alternatives have been chosen among the rest because they are the most extended ones.

5.1 BPEL and BPEL4People

We have chosen the tool provided by Oracle, Oracle BPEL Development Tool [9] to test BPEL implementation possibilities.

The initial idea was modeling the entire process “To Save Service Request” as an indivisible set, a “black box” to other business processes, as web services do. In this

process, we would only need to know its interface and address (information contained in wsdl file). To get our goal we would need to know all the parameters needed in the entire process at the beginning of it, the parameters needed for the first activity operation, the parameters needed for the second activity operations, and so on, and passing them to the process as input parameters. There is no way to know the value of some parameters needed in “Save Service Request”, such as Citizen ID, because they are being provided by another activity operation, “Search Citizen”. Using BPEL, we cannot ask the user about new parameters in the middle of the process because there is not any node similar to a user interaction. No questions are allowed. No interruptions are allowed. We find that it is not possible modeling user interactions in BPEL.

Further problem and other proposal analysis such as BPEL4People [10] take us to reconsider BPEL as modeling language but adding some changes to model user interactions. We have classified user interactions into two categories: those that require instant actions from users, called online interactions, and those that require actions from users but not now, maybe after one minute, hour, day or month, called offline interactions. Human billing approval is an offline interaction example. Offline interactions could be implemented using standard *Invoke*. BPEL *Invoke* is a web service call addressed to a service endpoint, or URL. The calling process waits for a callback message in a *Receive* or *Pick* activity, reporting that the human task has completed and returning result data. This *task management service* could also have state machines, assignment or escalation capabilities like human would do.

However, online interactions cannot be solved using task management. *We cannot wait an user answer for an undefined period of time* because the results do have a short life and maybe when user task has completed, the record (result) choose by user was deleted or updated by another. We cannot block records, as RDBMS does, while the system waits for a user answer in an undefined period.

5.2 WS-Coordination/WS-Transaction

We thought about the possibility of dividing the entire process into parts, modeling each part as an independent business process and coordinating the whole thing with WS-Coordination [11] or WS-Transaction [12]. However, possibilities given by these technologies are not much better.

WS-Coordination and WS-Transaction seem to be more suitable for distributed environments where orchestrating processes ensure operations sequence in order to complete successfully the whole transaction. Our process is not distributed; it does not run actions in several other systems that must be executed in a predefined sequence. In “*To Save Service Request*” process, user needs to take part in the process interacting with activity operations.

Moreover, similar problems to BPEL have arisen, for example, we do not resolve *everlasting wait* problem. Our business process model cannot be implemented with WS-Coordination/WS-Transaction.

5.3 “Inside Application Logic” Solution

Solution to this implementation problem is to develop business process logic “inside” web application logic, that is, web application knows exactly the next step and makes

direct invocations to web services (calling to external process). Moreover, web application controls the user interface logic. This solution is not free of inconveniences: strong coupling, business logic and presentations logic are assembled so changes in one will produce changes in the other; and low reusability, business logic cannot be reused or presented in other format, are its main objections.

6 Conclusions and Future Works

In this paper, we have presented an analysis of existing business process modeling alternatives to model those business processes with human-performed activities. We have followed a modeling method discussed on previous works that allows us to generate services composition model focusing on conceptual user services, those services the user really needs from the system.

The origin of this work was the difficulty to model human user interactions with most-extended technologies for business process composition, such as BPEL or WS-Coordination. BPEL does not accommodate human tasks in its specification although we have found a possible workaround it is only valid for some situations. WS-Coordination and WS-Transaction are specifications oriented to distributed environments where business processes do run on different systems with different allocations. Operations sequence, transaction atomicity and global consistency must be guaranteed in these situations.

Solution to this problem is implementing business logic inside application logic, although this solution arise strong coupling and low reusability disadvantages.

As future works, we propose on one hand, working on "inside web application logic" solution, merging business logic and presentation logic to find the improvements that could be implemented. On the other hand, working in new proposals for business process composition such as BPEL4People, which allows accommodating one of most common business process activity, human-performed activity.

References

1. Verner, L.: BPM The Promise and the Challenge. Queue of ACM, Vol. 2, No. 4. (2004) 82-91
2. Microsoft Corporation: XLANG: Web Services for Business Process Design. Available: http://www.gotdotnet.com/team/xml_wsspecs/xlang-c/default.htm [Accessed 21 June 2006] (2001)
3. Leymann, F.: Web Services Flow Language (WSFL 1.0). Available: <http://www-306.ibm.com/software/solutions/webservices/pdf/WSFL.pdf> [Accessed 21 June 2006] (2001)
4. Andrews, T., Curbera, F., Dholakia, H., Golan, Y., Klein, J., Leymann, F., Liu, K., Roller, D., Smith, D., Thatte, S., Trickovic, I., Weerawarana, S.: Business Process Execution Language for Web Services, Version 1.1 Specification. BEA Systems, IBM Corp., Microsoft Corp., SAP AG, Siebel Systems (2003)
5. De Castro, V., Marcos, E., López Sanz, M.: A Model Driven Method for Service Composition Modeling: A Case Study. Int. Journal of Web Engineering and Technology, (Accepted for publication) (2005)

6. Cáceres, P., Marcos, E., Vela, B.: A MDA-Based Approach for Web Information System Development. Proceedings of Workshop in Software Model Engineering Available: <http://www.metamodel.com/wisme-2003/> [accessed 7 May 2006] (2003)
7. Miller, J., Mukerji, J. (eds.), MDA, OMG Model Driven Architecture, Document number ormsc/2001-07-01, 2001. Available: <http://www.omg.com/mda> [accessed 7 May 2006].
8. Marcos, E., Cáceres, P., De Castro, V.: An approach for Navigation Model Construction from the Use Cases Model, Proceedings of 16th Conference on Advanced Information Systems Engineering, CAISE'04 FORUM. (2004) 83-92
9. Kodabackchian E.: Oracle BPEL Process Manager 2.0 (tutorials). Available: <http://otn.oracle.com/bpel> [accessed 29 June 2006]
10. Kloppmann M. et al: IBM Corporation. Available: <ftp://www6.software.ibm.com/software/developer/library/ws-bpel4people.pdf> [accessed 03 July 2006]
11. Cabrera F., Copeland G., Freund T., Klein J., Langworthy D., Orchard D., Shewchuk J., Storey T.: Web Services Coordination (WS-Coordination). BEA Systems, International Business Machines Corporation, Microsoft Corporation, Available: <http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnglobspec/html/ws-coordination.asp> [accessed 03 July 2006] (2002)
12. Cox W., Cabrera F., Copeland G., Freund T., Klein J., Storey T. Thatte S.: Web Services Transaction (WS-Transaction). BEA Systems, International Business Machines Corporation, Microsoft Corporation. Available: <http://dev2dev.bea.com/pub/a/2004/01/ws-transaction.html> [accessed 03 July 2006] (2002)

A Petri Net Based Approach for Reliability Prediction of Web Services

Duhang Zhong and Zhichang Qi

School of Computer Science, National University of Defense Technology
Changsha, 410073, P.R. China
zhong_duhang@hotmail.com

Abstract. Web service composition is a distributed model to construct new web service on top of existing primitive or other composite web services. Many research efforts have been made in web services composition and coordination. In such a scenario, an important issue is how to assess the degree of trustworthiness one can have about the web service composition, especially their performance and dependability characteristics. In this paper, we propose a petri net based approach to predict the reliability of web service composition. The first step of the approach involves the transformation of web service composition specification into Stochastic Petri Nets (SPN) model. The proposed transformation is built upon BPEL. From the SPN model, we can derive the reliability and performance measure of web service composition.

1 Introduction

Web services are emerging as a major technology for deploying automated interactions between distributed and heterogeneous applications. Various standards support this deployment including HTTP, XML, SOAP, WSDL [1], and UDDI [2]. Once deployed, web services provided by various organizations can be inter-connected in order to implement business process, leading to composite web services. For business process modeling different languages have been proposed, most of which are based on XML or graphical notations. Many such languages have recently emerged, including BPEL [3], BPML [4] or ebXML, these languages focus on tracking and executing collaborative business processes by business applications.

An important issue for business process built in this way is how to assess the degree of trustworthiness, especially their performance and dependability characteristics. In this paper we focus on reliability aspects, and propose an approach to predict the reliability of web service composition.

Stochastic Petri Nets (SPNs) [21] can be used to specify the problem in a concise fashion and the underlying Markov chain can then be generated automatically. In this paper, we propose the usage of stochastic petri nets as a solution to the problems of predicting the reliability of web service composition. The choice of petri nets was motivated by the following reasons: (a) Petri nets are a graphic notation with formal semantics, (b) the state of a petri net can

be modeled explicitly, (c) the availability of many analysis techniques for petri nets.

The remainder of this paper is organized as follows. Section 2 introduces, in a nutshell, both BPEL and stochastic petri net. In Section 3 we describe our reliability prediction method. Section 4 discusses the result of this mapping on an example BPEL process models found in [5]. Next, Section 5 discusses related works. Finally, we conclude this paper.

2 Preliminaries

2.1 BPEL

BPEL, also known as BPEL4WS, build on IBM's WSFL (Web Services Flow Language) and Microsoft's XLANG(Web Services for Business Process Design). It combines the features of a block structured process language (XLANG) with those of a graph-based process language (WSFL). BPEL is intended to describe a business process in two different ways: executable and abstract processes. An abstract process is a business protocol specifying the message exchange behavior between different parties without revealing the internal behavior of any of them. An executable process specifies the execution order between a number of constituent activities, the partners involved, the message exchanged between these partners, and the fault and exception handling mechanisms [11].

A composite service in BPEL is described in terms of a process. Each element in the process is called an activity. BPEL provides two kinds of activities: primitive activities and structured activities. Primitive activities perform simple operations such as receive (waiting for a message from an external partner), reply (reply a message to a partner), invoke (invoke a partner), assign (copying a value from one place to another), throw (generating a fault), terminate (stopping the entire process instance), wait (wait for a certain time), empty (do nothing).

To enable the representation of complex structures, a structured activity is used to define the order on the primitive activities. It can be nested with other structured activities. The set of structured activities includes: sequence (collection of activities to be performed sequentially), flow (specifying one or more activities to be performed concurrently), while (while loop), switch (selects one control path from a set of choices), pick (blocking and waiting for a suitable message). The most important structured activity is a scope. A scope is a means of explicitly packaging activities together such that they can share common fault handling and compensation routines. It consists of a set of optional fault handlers (exceptions can be handled during the execution of its enclosing scope), a single optional compensation handler (inverse some effects which happened during the execution of activities), and the primary activity of the scope which defines its behavior [11] [12] [13].

Structured activities can be nested. Given a set of activities contained within the same flow, the execution order can further be controlled through links. A link has a source activity and a target activity; the target activity may only start when the source activity has ended. With links control dependencies between concurrent activities can be expressed.

2.2 Stochastic Petri Nets

Petri Nets is a modeling formalism used for the analysis of a wide range of systems coming from different domains (e.g., distributed computing, flexible manufacturing, telecommunication, control systems, workflow management) and characterized by situations of concurrency, synchronization, causality and conflict [22]. Among the petri net based model types, we consider generalized stochastic petri nets (GSPNs) [23] and stochastic reward nets (SRNs)[21]:

Definition 1. A Generalized Stochastic Petri Nets (GSPN) is a 6-tuple $(P, T, F, W, M_0, \lambda)$. $P = \{p_1, p_2, \dots, p_k\}$ is a finite set of places. T is a finite set of transitions partitioned into two subsets: T_I (immediate) and T_D (timed) transitions, where transition $t \in T_D$ are associated with rate λ . $F \subseteq (P * T) \cup (T * P)$ is a set of arcs. $M_0 = \{m_{01}, m_{02}, \dots, m_{0k}\}$ is an initial marking. $W : T \rightarrow R$ is a function defined on the set of transitions. Timed transitions are associated with priority zero, whereas all other priority levels are reserved for immediate transitions. The immediate transitions are drawn as thin bars, while the timed transitions are drawn as rectangles.

The SRNs differ from the GSPNs in several key aspects [14]. The SRNs provide enabling functions, marking-dependent arc cardinalities, a more generalized approach to the specification of priorities, and the ability to decide in a marking-dependent fashion whether the firing time of a transition is exponentially distributed or null, often resulting in more compact nets. The models used in this paper are based on SRN, but simply we use term SPN instead.

3 Reliability Prediction Using Petri Nets

In this section we describe a method to predict BPEL composite web services reliability as a function of component web service reliability estimates. We transform the BPEL specification into a stochastic petri nets model, and annotate

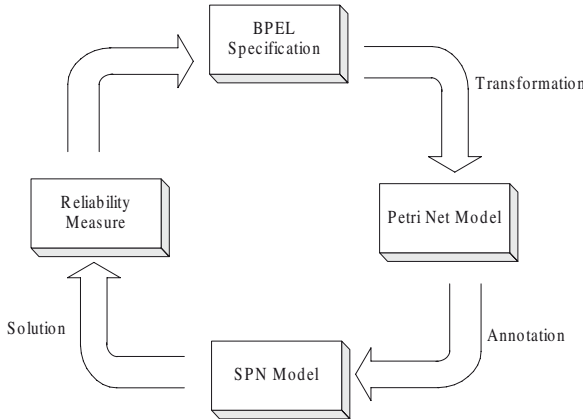


Fig. 1. Reliability prediction approach

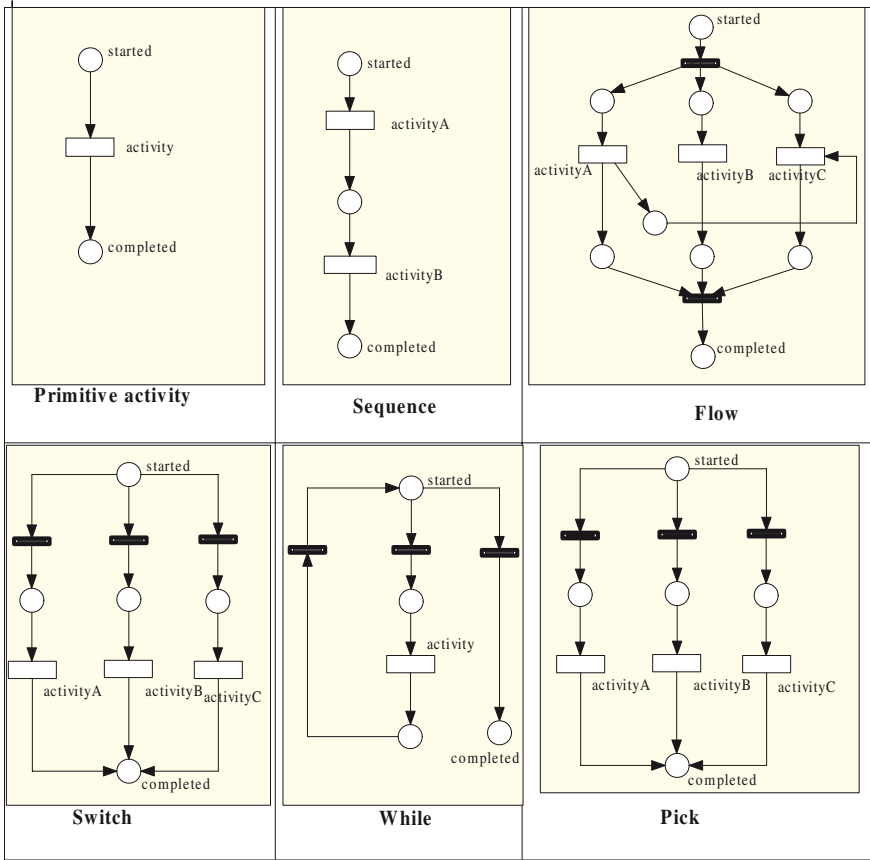


Fig. 2. Transformation of BPEL to Petri Nets

with probabilistic properties to predict the reliability of web service composition. The method is depicted in Fig. 1 as three major steps: (1) Transform BPEL specification into petri nets model, (2) annotation of the petri nets model, (3) solve the stochastic petri nets model and compute the reliability prediction.

3.1 Transformation from BPEL into Petri Nets Model

To represent BPEL using petri nets, basically we represent primitive activities with timed transitions. The control flow relations between activities specified in BPEL are captured by token firing rules and immediate transitions. The transformation details of primitive and structured activities into petri nets can be illustrated by these examples in Fig. 2.

3.2 Derivation of the SPN Model

In the second step, we annotate the petri nets model with the dependability attributes, and derive the SPN model of web service composition. There are three kinds of dependability attributes to be annotated:

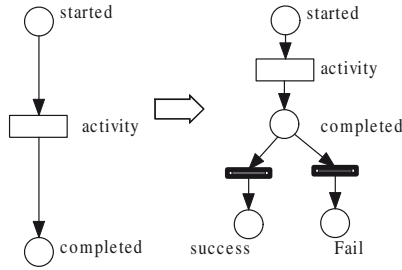


Fig. 3. Associate the failure behavior

- For every timed transition which represents the execution of a primitive activity, we annotate the execution time of the activity, which is assumed to be exponentially distributed with mean μ .
- For every immediate transition which represents the control structure relationship (eg. switch or while), we annotate to describe the weight assigned to the firing of enabled immediate transition t .
- In this paper, the reliability measure of a web service we use is the probability of its ability to successfully carry out its own task when it is invoked. To associate the failure behavior with the activities, we extend the petri net model transformed from BPEL in section 3.1. For each transition representing the execution of an activity offered by a web service, two immediate transitions added to represent the events that results produced by the activity are correct and incorrect respectively, and have weights (the reliability of the web service) r and $1-r$. This process is depicted as Fig. 3, Place "Fail" represents the failure of the BPEL composite web service.

3.3 Computing the Reliability Prediction

The last step is to solve the stochastic petri net model and compute the reliability prediction of web service composition. In this paper, we use the Stochastic Petri Net Package (SPNP) [15] [21] to computation of the reliability measures. SPNP is a versatile modeling tool for stochastic petri net model; it allows the specification of SPN models, the computation of steady-state, transient, cumulative, time-averaged, and up-to-absorption measures and sensitivities of these measures. The most powerful feature of SPNP is the ability to assign reward rates at the net level and subsequently compute the desired measures of the system being modeled [15]. Here we assign reward rate 1 to all markings in which there is no token in place "Fail"; all other markings are assigned a reward rate equal to zero. And the reliability of BPEL composite web service is the expected reward rate in steady state.

4 Examples

The following example shows how the structure of a BPEL process model is transformed into a stochastic petri nets model. Fig. 4 is the schematic illustration

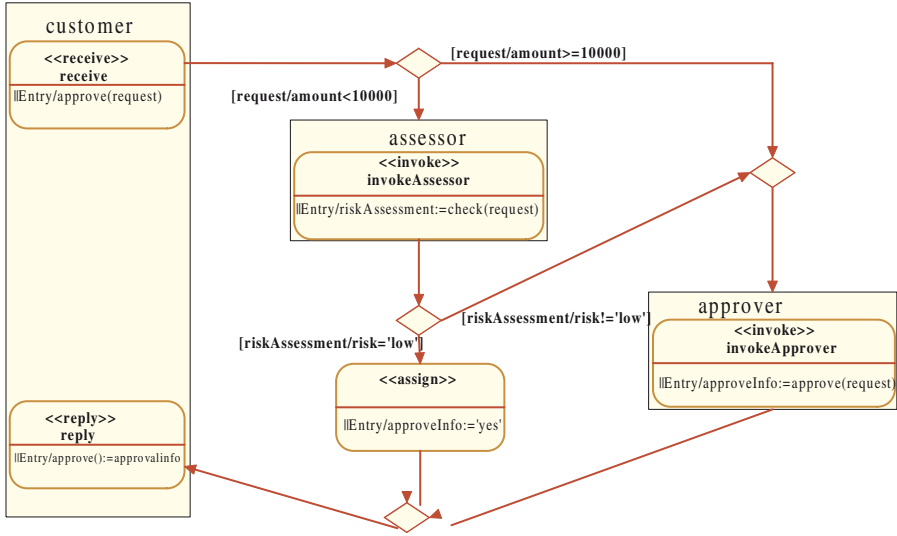


Fig. 4. Loan approval process

of the example taken from the section on structured activities of the BPEL 1.1 specification [5].

This example considers a simple loan approval web service that provides a port where customers can send their requests for loans. Customers of the services send their loan requests, including personal information and amount being requested. Using this information, the loan service runs a simple process that results in either a "loan approved" message or a "loan rejected" message. The approval decision can be reached in two different ways, depending on the amount requested and the risk associated with the requester. For low amounts (less than \$10,000) and low-risk individuals, approval is automatic. For high amounts or medium and high-risk individuals, approval is to be studied in greater detail. The corresponding stochastic petri nets model is depicted as Fig. 5.

In this example, the following parameters must be assigned a value before the SRN model can be evaluated:

- the reliability of each partner
- the probability weights of the immediate transitions
- the execution time of each primitive activity

We assume the values given in Table 1. Using the SPNP 6.0, we compute the reliability prediction for the loan approval process as $Rel = 0.948 = 94.8\%$

5 Related Work

Approaches to the reliability analysis of service- and component-based system have been already presented [6] [7] [8] [9]. According to the classification

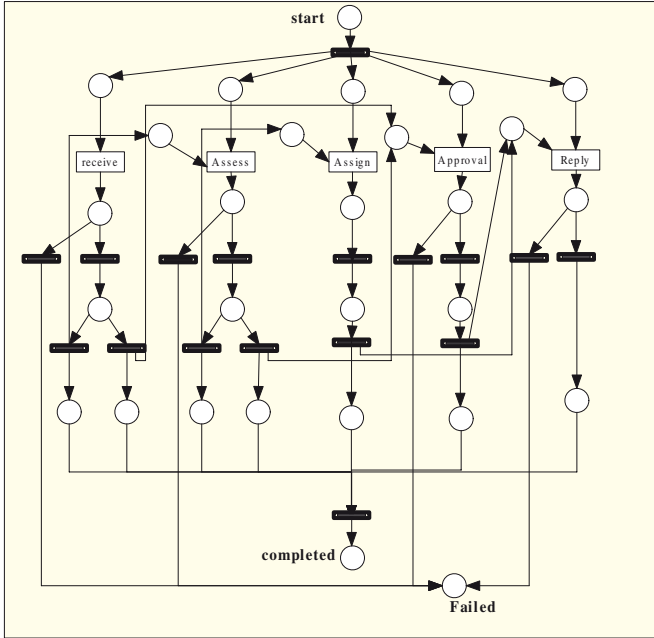


Fig. 5. The SPN model of loan approval process

Table 1. The parameters of loan approval process

Reliability	Value	Probability	Value	Execution time	Value
$R_{customer}$	0.98	$pr\{amounts \leq 10000\}$	0.40	$T_{receive}$	4
$R_{Assessor}$	0.99	$pr\{amounts > 10000\}$	0.60	T_{reply}	4
$R_{Approver}$	0.99	$pr\{risk = low\}$	0.30	T_{assign}	1
		$pr\{risk = high\}$	0.70	$T_{invoke_{Assessor}}$	10
				$T_{invoke_{Approver}}$	15

proposed by Goseva Popstojanova [10], they can be divided into two main categories: state-based approaches and path-based approaches. For the sake of brevity, we provide here a brief view of the approaches of greatest interest to the scope of this work.

State-based models [7] [16] use a control flow graph to represent the system architecture. In such models it is assumed that the transfer of control among the components can be modeled as a Markov chain, with further behavior of the system dependent only on the current state. The architecture of software has been modeled as a discrete time Markov Chain (DTMC), continuous time Markov Chain (CTMC), or a semi-Markov process (SMP). These can be further classified into absorbing and irreducible [10]. The former represents applications that operate on demand which software runs that correspond to terminating execution can be clearly identified. The latter is well suited for continuously

operating software applications, such that in real time control systems, where it is either difficult to determine what constitutes a run or there maybe very large number of such runs if it is assumed that each cycle consists a run.

Path-based models [17] compute the reliability of the system by enumerating possible execution paths of the program. The model used in their approach is the component dependency graph (CDG), this reliability analysis technique is specific for component based software whose analysis is strictly based on execution scenarios. A scenario is a set of component interactions triggered by specific input stimulus, and it is related to the concept of operations and run-types used in operational profiles[18].

In [9], Vincenzo Grassi present an approach to the reliability prediction of an assembly of services, that allows to take into account in an explicit and composition way the reliability characteristics of both the resources and interaction infrastructures used in the assembly. What distinguishes their approach is the exploitation of a "unified " service model that helps in modeling and analyzing different architectural alternatives, where the characteristic of both "high level" services and "low level" services are explicitly taken into consideration. Moreover, this work also point out the importance of considering the impact on reliability of service sharing.

In [8], Apostolos focused on the development of a principled methodology for the dependability analysis of composite Web Services. The first step involves a UML representation for the architecture specification of composite web services. The proposed representation is built upon BPEL and introduces necessary extensions to support the dependability analysis. The automated mapping of this extended UML models to traditional dependability analysis models such as Block Diagrams, Fault Trees and Markov models is the core of the methodology.

As pointed out by Jens Happe [19], most of the reliability analysis models are based on Markov models. A Markov model can be seen as a finite state machine, whose transitions are annotated with a probability of taking the transition from its source stats. These models can be appropriate when dealing with sequential systems. However, as soon as a concurrent or parallel software system (e.g. web service composition) has to be analyzed, which can hardly be expresses by finite state machines or the corresponding Markov model.

6 Conclusion

In this paper, we introduce an approach to predict the reliability of Web services composition. We present the transformation algorithms from BPEL, which is the de facto industry standard of Web services composition specification, to stochastic petri nets models. Using the model, we can compute the reliability prediction of the web service composition. The major contribution of this paper is a reliability prediction technique that takes into account the structure of BPEL specification and the concurrent nature of service composition. For future work, we will use our stochastic petri net model to give a more precise estimation of the reliability and performance of web service composition.

References

1. <http://www.w3.org/TR/wsdl>
2. F. Curbera, M. Duftler, R. Khalaf, W. Nagy, N. Mukhi, and S. Weerawarana. "Unraveling the web services web: An introduction to SOAP, WSDL, and UDDI", *IEEE Internet Computing*, 6(2):86-93, March 2002.
3. <http://www128.ibm.com/developerworks/library/ws-bpel/>
4. BPML.org. Business process modeling language. www.bpmi.org, 2002.
5. BEA, IBM, Microsoft, SAP AG, and Siebel Systems. Business process execution language for web services (version 1.1). <ftp://www6.software.ibm.com/software/developer/library/ws-bpel.pdf>, 2003.
6. R.H. Reussner, H.W.Schmidit, I.H.Poernomo "Reliability prediction for component-based software architectures" *Journal of Systems and Software*, no.66,2003.pp241-252.
7. W-L,Wang, Y.Wu, M-H Chen "An Architecture-based software reliability model" *Proc. IEEE Pacific Rim Int. Symposium on Dependable Computing*, Hong Kong China, Dec.1999.
8. Apostolos Zarras, Panos Vassiliadis, and Valerie Issarny, "Model-Driven Dependability Analysis of Web Services", In *Proceedings of the International Conference on Distributed Objects and Applications (DOA)*, LNCS3291,2004
9. Vincenzo Grassi, "Architecture-Based reliability Prediction for Service-Oriented Computing", *Architecting Dependable Systems III*, LNCS 3549,pp.279-299,2005.
10. K.Goseva-Popstojanova, A.P. Mathur, K.S.Trivedi "Comparision of architecture-based software reliability models" in *Proc. Of the 12th Int. Symposium on Software Reliability Engineering(ISSRE 2001)*, 2001.
11. H.M.W. Verbeek and W.M.P. van der Aalst. Analyzing BPEL Processes using Petri Nets. In D. Marinescu, editor, *Proceedings of the Second International Workshop on Applications of Petri Nets to Coordination, Workflow and Business Process Management*, pages 59-78. Florida International University, Miami, Florida, USA, 2005.
12. Sebastian Hinz, Karsten Schmidt, and Christian Stahl, "Transforming BPEL to Petri Nets.". *Proc. 3rd Int. Conf. on Business Process Management (BPM 2005)*, LNCS 3649, Nancy, France, 2005, pp. 220-235.
13. Humboldt-Universit At Zu, "Analyzing Web Service based Business Processes Axel Martens"2005
14. Gianfranco Ciardo, Jogesh K. Muppala and Krishor S. Trivedi, "Analyzing Concurrent and Fault-Tolerant Software using Stochastic Reward Nets", *Journal of Parallel and Distributed Computing*, Vol. 15, pp. 255-269, 1992.
15. C. Hirel, B. Tuffin, and K. S. Trivedi, SPNP : Stochastic Petri Nets. Version 6.0, in *Computer performance evaluation: Modelling tools and techniques; 11th International Conference; TOOLS 2000*, Schaumburg, Il., USA, B. Haverkort, H. Bohnenkamp, C. Smith(eds.), *Lecture Notes in Computer Science 1786*, Springer Verlag, 2000.
16. R.C.Cheung. A User-Oriented Software Reliability Model. In *IEEE Transactions on Software Engineering*, volume 6(2), pages 118-125. Mar.1980.
17. S.M.Yacoub, B.Cubic, and H.H.Ammar. Scenario-Based Reliability Analysis of Component-Based Software. In *Proc. Of the 10th ISSRE*, Boca Raton, FL,USA. IEEE, Nov.1999.
18. J.D.Musa, "Opeartional profiles in software reliability engineering.", *IEEE Software* 10(2) , 1993.

19. Jens Happe, Viktoria Firus, "Using Stochastic Petri Nets to Predict Quality of Services Attributes of Component-Based Software Architectures", the Tenth International Workshop on Component-Oriented Programming, Glasgow, Scotland (July 25-29, 2005)
20. Zhangxi Tan, Chuang Lin, Hao Yin, Ye Hong, and Guangxi Zhu, "Approximate Performance Analysis of web Services Flow Using Stochastic Petri Net", GCC 2004, LNCS 3251,pp.193-200, 2004.
21. <http://www.ee.duke.edu/chirel/research.html>.
22. Simona Bernardi, Phd Paper, "Building Stochastic Petri Net models for the verification of complex software systems", Torino.
23. Marsan A., Balbo G., Conte G., Donatelli S., Franceschinis G.: "Modeling with Generalized Stochastic Petri Nets", Wiley, Chichester, England, 1995.

Facilitating Ontology (Re)use by Means of a Categorization Framework

Peter De Baer, Koen Kerremans, and Rita Temmerman

Erasmushogeschool Brussel, Centrum voor Vaktaal en Communicatie,
Trierstraat 84, 1040 Brussels, Belgium
{Peter.De.Baer, Koen.Kerremans, Rita.Temmerman}@ehb.be

Abstract. Ontologies as means for conceptualizing and structuring domain knowledge within a community of interest are seen as a key to realize the Semantic Web vision. However, the decentralized nature of the Web makes achieving consensus across communities difficult, thus, hampering efficient knowledge sharing between them. To address this problem of heterogeneity we propose a Categorization Framework (CF) that makes it possible to use (multilingual) terminology to specify concepts and concept relations in domain ontologies. Such CF could describe the meaning of concepts and concept relations by means of terminological information and external references. We believe that such (multilingual) ontology description could enhance the (re)usability and facilitate the coordination¹ of domain ontologies.

Keywords: categorization framework, multilingual terminology, ontology coordination and ontology description.

1 Facilitating Ontology (Re)use

Ontologies as means for conceptualizing and structuring domain knowledge within a community of interest are seen as a key to realize the Semantic Web vision. However, the decentralized nature of the Web makes achieving consensus across communities difficult, thus, hampering efficient knowledge sharing between them.

In this paper we describe a categorization framework (CF), which could be used to address the problem of heterogeneity of domain ontologies at the terminological level [2]. The CF has been the result of research in the fields of terminology engineering and knowledge engineering and is part of the Termontography methodology, a multidisciplinary approach in which theories and methods for multilingual terminological analyses [12] are combined with methods and guidelines for ontology engineering. A clear distinction is made between conceptual modelling at a language-independent level and a language-specific analysis of units of understanding [7].

A CF can be seen as an ontological structure, i.e. concepts and concept relations, enriched by terminological information. As will be shown, the CF could be used to identify and specify concepts and concept relations by means of terminology from

¹ Ontology coordination: broadest term that applies whenever knowledge from two or more ontologies must be used at the same time in a meaningful way [2].

(specialized) natural language. By doing so, we believe that a CF could enhance the accessibility and facilitate the coordination (i.e. mapping, alignment, merging) [9] of domain ontologies. As a result the (re)usability of domain ontologies may increase.

In section 2, we illustrate the difficulties of ontology coordination with regard to identification. In section 3, we show how terminology may effectively be used to identify concepts and concept relations. In section 4, we show how ontology coordination is further complicated if the source ontologies have been developed in different linguistic and/or cultural settings. Section 5 describes a CF and how it can overcome the complexity of coordinating or harmonizing ontologies. In section 6, we will discuss how a CF can be used as an application ontology. Finally, in section 7, we conclude.

2 The Need for Identification of Concepts and Concept Relations

Let us conduct a small thought experiment. Suppose we have two different domain ontologies we want to coordinate with one another (*see figure 1*). The first domain ontology O1 has four concepts C1, C2, C3 and C4, while the second domain ontology O2 has three concepts c1, c2 and c3. Between the concepts of O1 exist the following directed concept relations²: C1->(R1)->C2, C1->(R1)->C3, C1->(R1)->C4, C3->(R2)->C2 and C3->(R2)->C4. Between the concepts of O2 exist two directed concept relations c1->(r1)->c2 and c2->(r2)->c3. Let us now try to merge O1 and O2 into a third encompassing domain ontology EO. In order to succeed, there is a need for a common vocabulary, which transcends both O1 and O2 and which allows us to identify the identical concepts and directed concept relations. In case of domain ontologies, this common vocabulary may be established through the use of words or expressions (i.e. terminology) derived from (specialized) communicative settings.

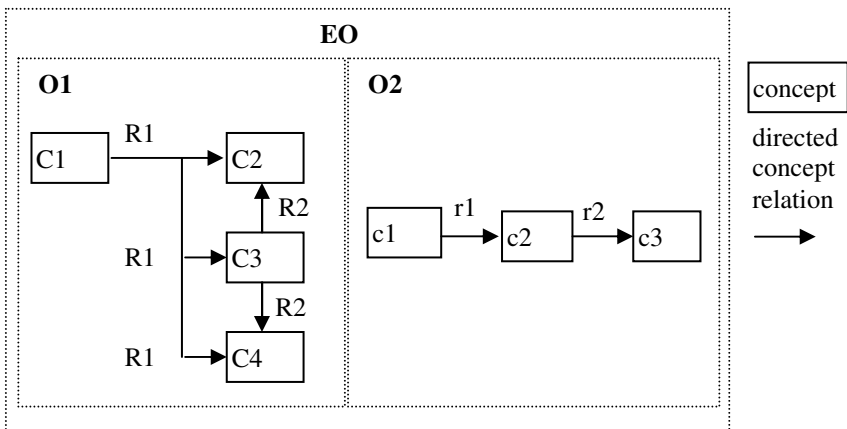


Fig. 1. Merging domain ontologies O1 and O2 into EO

² We notate a directed concept relation dcr1 between to concepts c1 and c2 as c1->(dcr1)->c2.

3 Using Terminology to Identify Concepts and Concept Relations

We continue our thought experiment (*see section 2*) to show how terminology may effectively be used to identify the concepts and concept relations in O1 and O2, and to be able to merge the two domain ontologies into EO. We add English terminology to label the concepts of both O1 and O2: C1=arm, C2="lower arm", C3=wrists, C4=hand, c1=hand, c2=finger and c3=thumb. Next, we add English terminology to the directed concept relations of both O1 and O2: R1="has part", R2="is connected to", r1="has part" and r2="has subordinate concept". The naming information allows us to reduce EO as we see that C4=c1 and R1=r1. As a result, EO may now be represented as follows (*see figure 2*): arm->(has part)->"lower arm", arm->(has part)->wrist, arm->(has part)->hand, wrist->(is connected to)->"lower arm", wrist->(is connected to)->hand, hand->(has part)->finger and finger->(has subordinate concept)->thumb.

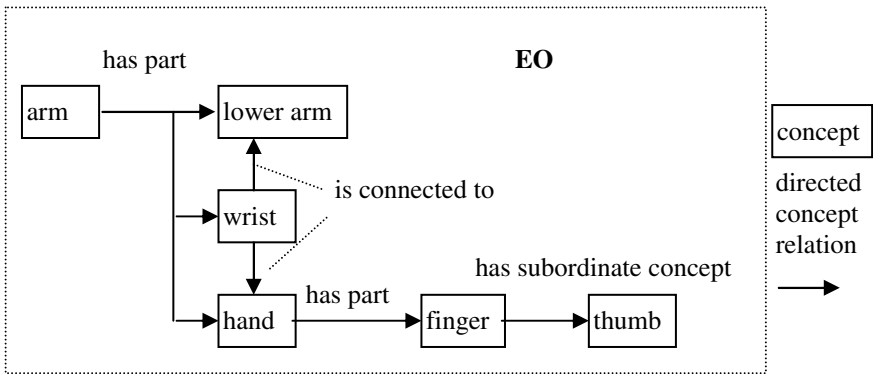


Fig. 2. Identification of concepts and concept relations by means of terminology

4 Multilingual, Multicultural and/or Multisituational Settings

The task of ontology coordination is further complicated if the source ontologies have been developed in different linguistic and/or cultural settings. Suppose for example that O1 was developed in France using French terminology to identify the concepts and concept relations, while O2 was developed in the UK using English terminology. To align both O1 and O2 we could add terminology to O1 and O2 (*see table 1*) or build a new encompassing ontology EO (*see table 2*) using multilingual terminology to identify concepts and concept relations.

From the previous example we might conclude that building multilingual ontologies and using multilingual terminology to identify concepts and concept relations is **the** solution to support ontology coordination. However, one needs to take into account the many interpretation problems that may arise due to the existence of polysemous words or homographs. The English word *crane*, for instance, may refer to a machine for hoisting and moving heavy objects by means of cables attached to a movable boom, as well as any of various large wading birds of the family *Gruidae* [1].

Table 1. Adding terminology to O1 and O2

	O1		O2	
Concept	French	English	English	French
C4	main	hand		
c1			hand	main

Table 2. Building an encompassing ontology EO

	O1	O2	EO	
Concept	French	English	French	English
C4	<i>main</i>			
c1		hand		
Ca			<i>main</i>	hand

Apart from polysemy, semantic vagueness could be caused by cultural and situational differences regarding the meaning of an expression. The expression *nursing home* for instance will have different (legal) characteristics in different English speaking communities governed by different legal systems (e.g. England, Irish Republic, USA, etc.). A *nursing home* has thus a culture-specific meaning besides its broad general meaning.

In order to accurately identify concepts and concept relations by means of terminology, it is necessary to further specify the context of the terminology used. A simple form of such contextualization can already be seen in table 2. Terms in this multilingual setting should not be considered as mere text strings, but rather as combinations of a text string and a language. The French word *main*, for instance, means *hand*, while the English word *main* could refer to a *chief or largest part* [1].

In the next section we will describe a CF that allows us to better specify the context, i.e. the meaning, of terms.

5 Categorization Framework

A *categorization framework* consists of the following items: *term*, *category*, *meta category*, *term meaning*, *attribute*, *property*, *bi-directional relation* and *bi-directional relation instance*.

- A **term** is a text string that is classified by one or more categories. Each term must be classified by only one language category.

For instance, the text string *main* classified by the French language category fr^3 constitutes the French term *main*.

³ Following ISO 639-1: codes for the identification of languages.

- A **category** belongs to a single superordinate concept we call a *meta category* and can be used to classify items, for instance terms.

The English language category *en*, for instance, belongs to the meta category *language*.

- A **meta category** specifies the superordinate concept of all the categories that belong to it.

If we add the category *fr*, for instance, to the meta category *language*, this category represents the French language.

By means of these three concepts *meta category*, *category* and *term* we are able to contextualize terminology. For example, we could create a meta category *regional language* and add the category *en-UK*⁴ to this meta category. This category *en-UK* could then be used to classify the English term *aubergine* as a British English term (the common American English term for this fruit is *eggplant*).

Although we now have a simple mechanism to add context to terminology, there are still some issues to resolve. Our goal is to use terminology to specify concepts and concept relations. We already learned that a text string is insufficient to refer to a certain meaning. The text string *main*, for instance, might refer to a *hand* when interpreted in French or a *chief or largest part* [1] when interpreted in English. For this reason we introduced the item *term* that is a combination of a language and a text string. This does not suffice, however, to solve the problem of polysemy and semantic vagueness. The English term *bow*, for instance, might refer to *the front part of a ship*, *to bend*, *a weapon*, etc. To resolve this semantic ambiguity, we relate the closest superordinate concept to the concept we want to refer to. For instance, if we want to indicate the concept *weapon that shoots arrows*, we relate the superordinate concept *weapon* with the English term *bow*. To implement this classification method in a generic manner, we use a meta category to represent the superordinate concept and a category to represent the concept. The previous three meanings of *bow* could be implemented as described in table 3.

Table 3. Three meanings of the English word *bow*

Meta category	Category	Category description
part of a ship	bow	front part of a ship
verb	bow	to bend
weapon	bow	a weapon that shoots arrows

We should note that the identification of meta categories and categories here is no longer by means of *terms*. Instead, we use the concept *term meaning* that refers to a specific meaning.

- A **term meaning** is an item with both a reference to a *term* and a *meaning*, i.e. (meta) category. *Term meanings* may be added to a meta category and/or category.

⁴ Following RFC 3066: codes for the identification of (regional) languages.

With this definition a *term meaning* now refers to a certain meaning, while it still can be represented by a single term. The *term meaning* that references the English term *joint* and the category *marijuana cigarette* with meta category *cigarette*, for instance, clearly describes this *term meaning* as a marijuana cigarette.

We should also note that a meta category might be a category itself. Consider for instance the category *serpent* with meta category *wind instrument*, this meta category could be a category itself with meta category *musical instrument* [6]. To practically implement this, we further extend the CF with a *meaning* item.

- A **meaning** is the underlying item of a meta category and/or category. A meaning has a list of *term meanings* and may have references to both a meta category and a category.

The term meaning *wind instrument*, for instance, could reference a certain *meaning* with references to the meta category *wind instrument* and the category *wind instrument*.

By adding *term meanings* to a (meta) category using the underlying *meaning* item, the meaning of the (meta) category will be specified.

Figure 3 summarizes the CF items defined in this section and compares them with better-known notions from semiotic theory.

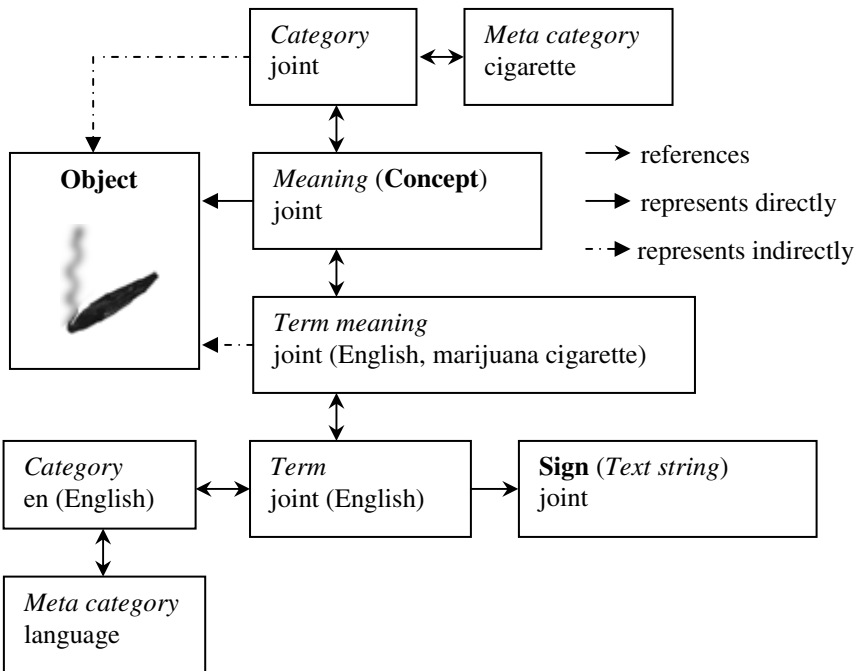


Fig. 3. The defined items (*italic*) in relation to the semiotic notions (*bold*) [13]

The sign *joint* in figure 3 is simply a text string. The English term *joint* that references this sign would be classified by the English language category *en*. The *term meaning*

that references both the English term *joint* and the category *marijuana cigarette* now clearly represents a specific *meaning*. This *meaning* directly represents the intended object. Consequently, the *term meaning* indirectly represents the same object.

By now we are able to identify concepts, i.e. categories, by means of terminology (using *term meanings*). In a similar way we could identify concept relations if we add the meta category *relation* and add for each concept relation a specific category to this meta category. For instance the category *has part* could represent the (physical) whole-to-part relation.

- Such classification structure of meta categories and categories is what constitutes a **categorization framework**.

To allow such CF to be useful when aligning domain ontologies, the choice of the meta category of each category is important. The meta category should be the closest superordinate concept of a category. By doing so, we should be able to implement a sufficiently fine-grained classification structure.

We will now demonstrate how a CF can be aligned with the domain ontology EO described in sections 2 and 3. We start by creating the necessary meta categories and categories to represent the concepts in EO i.e. *arm, lower arm, wrist, hand, finger and thumb*. The superordinate concept of these concepts is *human body part*, so we create this meta category. To this meta category, we then add each concept as a category. Note that each combination of the meta category and a category uniquely identifies a concept.

The next step is to create the meta category *relation* and to add the concept relations of EO, i.e. *has subordinate concept, has part and is connected to*, as categories to this meta category. Again, the combination of the meta category and a category uniquely identifies a concept relation, at least in so far as the terminology itself clearly identifies the meaning of the concept relation.

To further clarify the intended meaning of a concept relation (or concept), we allow properties to be added to CF items (*categories* and *term meanings*, for instance). To implement properties we introduce the items *attribute* and *property*.

- An **attribute** is implemented as a category with meta category *attribute*. Each attribute should refer to a certain value type. The list of possible value types depends upon the specific implementation of the CF. The value type *text string* is a minimum requirement. The value types *URI*⁵, *URN*⁶ and *URL*⁷ have been proven very useful too.

For example, the attribute *description* with value type *text string* could be used to describe a *term meaning*, while the attribute *extra information* with value type *URL* could be used to refer to a web page with extra information about a category.

- A **property** references an attribute and a value of a certain value type. Properties may be added to each CF item i.e. *meta category, category, term, term meaning, attribute, property, bi-directional relation* and *bi-directional relation instance*.

For example, we could add a property with attribute *extra information* and value <http://en.wikipedia.org/wiki/Finger> to the category *finger* [5].

⁵ Uniform Resource Identifier.

⁶ Uniform Resource Name.

⁷ Uniform Resource Locator.

Using properties we can accurately describe *term meanings* and since *term meanings* refer to concepts, the same goes for the concepts in the CF. Furthermore, *term meanings* enable us to use terminology to identify concepts. In combination with the use of (meta) categories for classification, this allows us to add context to terminology for disambiguation and description purposes. We could, for example, describe the legal differences of the concept *maison de repos* in Belgium and France. In both legal systems the same term may be used to refer to a concept with a similar general meaning, however legal differences exist and must be described. To do so, we could add two *term meanings* with references to the category *rest home* and to the French term *maison de repos*. Using two categories *Belgium* and *France* with meta category *legal system* we could classify both *term meanings*. Properties with descriptive legal information could then be added to each *term meaning*.

Properties could also be used to refer to concepts in external domain ontologies by means of URIs. A concept in the CF could thus be linked to a concept in another domain ontology, which makes it possible to align the CF with other domain ontologies. We believe that such aligned CF could facilitate the process of ontology coordination since identification of the concepts and concept relations would be easier. In fact such aligned CF could be seen as an ontology itself, at least without regarding specific conceptual information like properties, characteristics, etc. The (automatic) coordination of domain ontologies would of course still require specialized tools since different ontology formats exist. For this reason, we believe it is important that the CF could be used in a generic manner by different tools. We therefore designed the CF as an application ontology, i.e. an extendible ontology structure that could be used by applications to control which information and how information should be displayed.

6 The CF as an Application Ontology

In the previous section we showed how the CF may be extended by adding (meta) categories. These (meta) categories can be addressed using terminology. Consequently, applications could use the CF as an application ontology. For example, a multilingual ontology viewer could use the language category *fr* to retrieve and display the French terminology of the concepts and concept relations.

To efficiently manage (meta) categories, the use of relations must still be explained. We should consider the fact that a relationship between two CF items has two directions, mostly with a different meaning. We therefore introduce the item *bi-directional relation* to specify a bi-directional relation between two CF items.

- A **bi-directional relation** references at least one relation and at most two relations.

As we already described in section 5, a relation can be created by adding a category with meta category *relation*. We may add the categories *is part of* and *has part*, for example, to the meta category *relation*. A logical *bi-directional relation* should reference both these two opposing categories. Let us notate this bi-directional relation as *((is part of), (has part))*.

- A **bi-directional relation instance** references a *bi-directional relation* and two *meta categories*, *categories* or *term meanings*. Since the direction of a *bi-directional relation* is usually relevant, the *bi-directional relation instance*

makes a distinction between the source and the target item. *Bi-directional relation instances* can be created between *meta categories*, *categories* and *term meanings*.

The *bi-directional relation instance* that references the bi-directional relation (*is part of*), (*has part*)), the source category *lower arm* and the target category *arm* would indicate that a *lower arm* is part of an *arm*.

Using *bi-directional relation instances* the CF can be structured in a flexible, yet generic manner. A hierarchical structure could be created, for example, on the list of meta categories. By doing so, this meta category hierarchy can be used to browse through the CF. Since each meta category may have multiple parents, multiple entry points can be provided to expose the underlying categories. The hierarchy of meta categories then has a function comparable to that of topics in a topic map⁸.

The hierarchical generic-specific relation should be implemented by means of the meta category specification of a category. Although a concept can thus have only one direct superordinate concept, different aspects of this concept could still be specified using categories to classify the concept. The concept *knee* with superordinate concept *human body part*, for instance, could be classified by both the categories *medicine* and *biology* with meta category *subject field*.

Using the described CF items and terminology, specialized applications could use the CF as an application ontology. For example, to display a French whole-to-part concept hierarchy the application could browse the list of *bi-directional relation instances*, select the *bi-directional relation instances* referencing the (*is part of*), (*has part*)) bi-directional relation and use the source and target categories to structure the concept tree. For each concept, the application should retrieve the list of *term meanings*. The *term meaning* that references a term with language category 'fr' should be used to represent the concept or if no French *term meaning* is available the application could use an alternative language.

7 Conclusions and Future Work

In this paper we describe a categorization framework (CF) that could be used to facilitate ontology (re)use. More specifically, (multilingual) terminological information could be used to identify and describe concepts and concept relations in domain ontologies. By doing so, the CF could improve the accessibility and facilitate the coordination of domain ontologies. For this purpose, a CF can also be used as an application ontology in a wide range of applications like knowledge engineering tools [3, 10, 11], terminology engineering tools [8], knowledge discovery tools [4], etc.

In the PoCeHRMOM project⁹ the CF has been set up to combine information about competencies and occupations from several existing multilingual knowledge resources. Based on this work, we believe that the CF is scalable since categories can be used to classify concepts, concept relations and properties. Thus sub-ontologies may

⁸ Topic maps are an ISO standard for the representation and interchange of knowledge, with an emphasis on the findability of information. The standard is formally known as ISO/IEC 13250:2003 [1].

⁹ <http://cvc.ehb.be/Projects.htm#PoCeHRMOM>

be filtered out by means of subject categories. However, to efficiently manage large CFs, more specialized collaborative software tools will still be needed.

In the future we want to develop collaborative CF engineering tools to integrate more existing information resources by means of a CF.

Acknowledgment. This research is part of the PoCeHRMOM-project, sponsored by IWT-TETRA and seven Flemish SMEs: Actonomy, Ascento, Co:Inpetto, Jobs & Careers, Linking, Synergetics and The Profile Group.

References

1. Answers Corporation: Online Encyclopedia, Thesaurus, Dictionary definitions and more, <http://www.answers.com/> (retrieved May 26, 2006)
2. Bouquet, P., Ehrig, M., Euzenat, J., Franconi, E., Hitzler, P., Krötzsch, M., Serafini, L., Stamou, G., Sure, Y., Tessaris, S.: D2.2.1 Specification of a common framework for characterizing alignment, KWEB EU-IST-2004-507482 (2005)
3. Buitelaar, P., Olejnik, D., Sintek, M.: OntoLT: A protégé plug-in for ontology extraction from text. In Proceedings of the International Semantic Web Conference (ISWC) (2003)
4. De Baer, P., Kerremans, K., Temmerman, R.: Bridging Communication Gaps between Legal Experts in Multilingual Europe: Discussion of a Tool for Exploring Terminological and Legal Knowledge Resources. Proceedings of the Euralex conference, Turin, 2006 (forthcoming)
5. Hepp, M., Bachlechner, D., Siorpaes, K.: Harvesting Wiki Consensus - Using Wikipedia Entries as Ontology Elements, Proceedings of the 1st Workshop: SemWiki2006 - From Wiki to Semantics, co-located with the 3rd Annual European Semantic Web Conference (ESWC 2006), June 12, 2006, Budva, Montenegro (forthcoming)
6. Jenkins, K.: Thesaurus of Musical Instruments, <http://www.alteriseculo.com/instruments/> (retrieved May 26, 2006)
7. Kerremans, K., Temmerman, R.: Towards Multilingual, Termontological Support in Ontology Engineering, Proceedings Workshop on Terminology, Ontology and Knowledge representation, Lyon, France (2004)
8. Kerremans, K., Temmerman, R., Tummers, J.: Discussion on the Requirements for a Workbench supporting Termontography. Proceedings of the Euralex conference, Lorient, France (2004)
9. Kotis, K., Vouros, G.A., Stergiou, K.: Towards Automatic Merging of Domain Ontologies: The HCONE-merge approach, Journal of Web Semantics (2005)
10. Noy, N.F., Musen, M.A.: The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping, International Journal of Human-Computer Studies, 59/6 pp. 983-1024 (2003)
11. Pazienza, M.T., Stellato, A.: An open and scalable framework for enriching ontologies with natural language content, The 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'06), special session on Ontology & Text Ancecy, France, June 27-30, 2006 (forthcoming)
12. Temmerman, R.: Towards New Ways of Terminology Description. The sociocognitive approach, Amsterdam/Philadelphia: John Benjamins (2000)
13. Trippel, T.: The terminologists' concept system, <http://www.spectrum.uni-bielefeld.de/~ttrippel/terminology/node13.html> (retrieved May 26, 2006)

Agent-Grid Integration Ontology

Frederic Duvert, Clement Jonquet, Pascal Dugenie, and Stefano A. Cerri

LIRMM, CNRS & University Montpellier 2
161 Rue Ada, 34392 Montpellier Cedex 5, France
{duvert, jonquet, dugenie, cerri}@lirmm.fr

Abstract. The integration of GRID and MAS (Multi-Agents Systems) is an active research topic. We have recently proposed the Agent-Grid Integration Language, to describe a service-based integration of GRID and MAS models. However, the complexity of the mutual integration aspects leads us to define a rigorous way to formalize the key concepts, their relations and the integration rules by means of an ontology. With this ontology, we can describe the elements and their composition that occur in various service exchange scenarios with agent on the Grid. The ontology could be used both to model the behaviour of GRID-MAS integrated systems and to check the consistency of these systems and their instances. A concrete scenario is illustrated.

1 Introduction

GRID and MAS (Multi-Agent Systems) are two kinds of distributed systems. Yet, the motivations are different. GRID focuses on a reliable and secure resource-sharing infrastructure, whereas MAS focus on flexible and autonomous collaboration and problem solving in uncertain and dynamic open environments. The GRID and MAS communities believe in the potential of GRID and MAS to enhance each other because these models have developed significant complementarities [1]. In [2,3] we explain why these two domains join with the concept of service:

- GRID and MAS have each developed a service oriented behaviour, therefore the concept of service may represent a common integration;
- New needs in service exchange scenarios are clearly highlighted (dynamicity, composition, conversation based, user-centred behaviour, business processes, semantics, etc.) [4] and may be met by integrating GRID and MAS complementarities.

One of the crucial explorations concerns the substitution by an agent-oriented kernel of the current object-oriented kernel of services available in GRID. In [2,3] we propose a model for GRID-MAS integrated systems which considers that services are exchanged (i.e., provided and used) by *agents* through *GRID* mechanisms and infrastructure. In this model, concepts, relations between them and rules of these systems are semantically described by a set-theory formalization and a common graphical description language, called *Agent-Grid Integration Language* (AGIL).

In this paper, we formalize, by means of an ontology the GRID-MAS integration model proposed by AGIL. This ontology, called *Agent-Grid Integration Ontology* (AGIO), describes the semantics of GRID-MAS integrated system elements as well

as the behaviour of GRID-MAS integrated systems. AGIO's GRID concepts are directly influenced by OGSA (Open Grid Service Architecture) [5] and AGIO's MAS concepts are influenced by different approaches in MAS, such as the STROBE model [6], the Agent-Group-Role (AGR) model [7], Belief-Desire-Intention (BDI) architectures, Foundation for Intelligent Physical Agents (FIPA) agents, or other elements of the MAS literature [8,9]. AGIO describes the elements implied in service-oriented interaction between agents. It is actually a meta-description, allowing agents to agree on what they are, what is a service, a host, etc.

2 GRID-MAS Integrated Model

Service-based integration of GRID and MAS models. The concept of service is clearly at the intersection of the GRID and MAS domains. GRID is said to be the first distributed architecture (and infrastructure) really developed in a service-oriented perspective: Grid services are compliant Web services, based on the dynamic allocation of virtualized resources to an instantiated service [5]. Whereas Web services have instances that are stateless and persistent, Grid service instances can be either stateful or stateless, and can be either transient or persistent. A stateful service has an internal state that persists over multiple interactions. For a recent precise overview of Grid service concepts and standardization, see for example [10].

On the other hand, agents are said to be autonomous, intelligent and interactive entities who may use and provide services (in the sense of particular problem-solving capabilities) [8,9]. Actually they have many interesting characteristics for service exchange: they are reactive, efficient, adaptive, they know about themselves, they have a memory and a persistent state, they are able to have conversation, work collaboratively, negotiate, learn and reason to evolve, deal with semantics associated to concepts by processing ontologies, etc. MAS and service-oriented computing recently turned to one another considering the important abilities of agents for providing and using dynamic composed/composite services, semantic services, business processes, etc. [4].

Key GRID and MAS concepts and their integration. GRID is a resource-sharing system. Grid resources are contributed by *hosts*. A host is either a direct association between a *computing resource* and a *storage resource* or a *host coupling*. The sharing of these resources is implemented by the virtualization and the reification of these resources in *service containers*. A *Grid service* is included in a hosting environment in order to exist and to evolve with their own private contexts (i.e., set of resources). This is the role of the service container which is the reification of a portion of the virtualized resource available in a secure and reliable manner. A service container contains several types of services. A service may instantiate another service in the same or different service container. Each service is identified by a *handle*. Since a container is a particular kind of service, it is created either through the use of a service factory or by the direct core GRID functionality. A service container is allocated to (and created for) one and only one group of *agents*,¹ called a *Virtual Organization (VO)*.² Each agent may belong to

¹ The term agent is used to uniformly denote artificial agent, human agent and Grid user.

² The term VO unifies the concept of VO in GRID and the concept of group in MAS.

several VOs. The relation between VO members and Grid services in the associated container is embodied by a *Community Authorization Service* (CAS) which formalizes the VO-dedicated policies of service by members. In order to participate in GRID, hosts and agents must hold a *X509 certificate* signed by a special authority.

An *agent* possesses both intelligent and functional abilities. These are represented respectively by the agent *brain* and *body*. The brain is composed of a set of rules and algorithms (e.g., machine learning) that give to the agent learning and reasoning skills. It also contains the agent knowledge, objectives, and mental states (e.g., BDI). The body is composed of a set of *capabilities* which correspond to the agent's capacity or ability to do something, i.e., to perform some task. These capabilities may be interfaced as Grid services in the service container that belongs to a VO an agent is a member of. In the agent's body, these capabilities may be executed in a particular context called a *cognitive environment*.³ A cognitive environment contains several capacities. An agent may have several cognitive environments which correspond to the different conversation contexts and languages it develops by interaction with other agents. These interactions can be for example service exchanges i.e., situations where agents use the service another agent provides.

The GRID-MAS integrated model is illustrated in Fig. 1. We sum-up here the two main underlying ideas:

- The representation of agent capabilities as Grid services in a service container, i.e., viewing Grid service as an 'allocated interface' of an agent capability by substituting the object-oriented kernel of Web/Grid services with an agent oriented one;
- The assimilation of the service instantiation mechanism – fundamental in GRID as it allows Grid services to be stateful and dynamic – with the dedicated cognitive environment instantiation mechanism – fundamental in STROBE as it allows one agent to dedicate to another one a conversation context.

3 GRID-MAS Integration Ontology

3.1 Ontology Modelling

When formalized and computerized, shared knowledge can serve as the basis for better understanding among agents. In recent years, the representation of such shared knowledge has largely been implemented by ontologies i.e., formal, computerized conceptualization of the notions, properties and relationships in a domain [11]. Nowadays, ontologies are used by agents each time a semantic description is needed. Ontologies are composed of *concepts*, *relations* and *instances*. For example, if you want to define a car, you should say: 'a car is a transportation object, with four wheels, and you need a licence to drive it. MyCar is a car.' 'Car' is a concept, 'is a' is a relation, and 'MyCar' is an

³ Conversations and their states are represented in the STROBE model [6] by cognitive environments. We do not detail this aspect here. In other agent architectures, cognitive environments may simply be viewed as conversation contexts. Our GRID-MAS integrated model, was influenced by the STROBE model as it is a communication and representation agent model developed in a service perspective that fits well Grid service mechanisms such as for example the concept of instantiation.

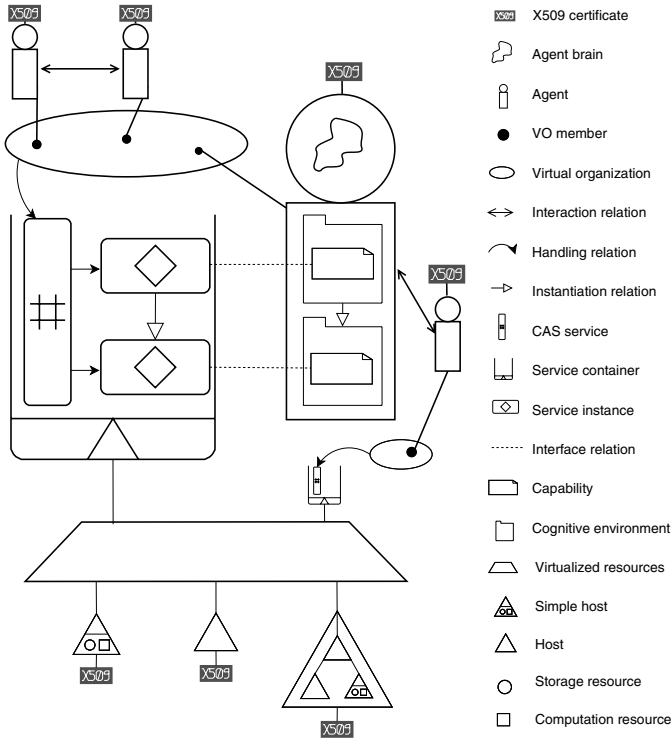


Fig. 1. The GRID-MAS integrated model described in AGIL's graphical description language

instance. Ontologies allow also to define *rules*. A rule is a constraint on a set of relations, that is not directly expressed by the relations. In the rest of the paper, we use the following writing conventions: *Concept*, *Relation* and *INSTANCE* e.g., *MYCAR is a Car*.

A common standard for representing ontology is the Web Ontology Language (OWL) [12]. Ontologies editors free the designer from writing directly XML/OWL code. They automatically generate it, allowing ontologies to be both understandable and processable by human and artificial agents. The Protégé knowledge-modelling environment (<http://protege.stanford.edu>) is a de facto standard tool that supports single and multi-user construction of ontologies. It enable designers to express rules by means of the Semantic Web Rule Language (SWRL) [13].

3.2 Agent-Grid Integration Ontology

GRID and MAS key concepts simplified and briefly summarized in Sect. 2 are translated into an ontology. Concepts are presented in Table 1 and relations in Table 2. The integration rules are expressed by means of SWRL rules. AGIO is a procedural ontology that describes the different elements of GRID-MAS integrated systems and their interactions. Adding an instance of a GRID-MAS integrated system to AGIO would make it evolve as a descriptive ontology.

Table 1. AGIO's concepts and inheritance

CONCEPT	INHERITS	CONCEPT	INHERITS
Resource	Thing	ServiceContainer	Service
ComputingResource	Resource	CAS	Service
StorageResource	Resource	NormalService	Service
Host	Thing	Agent	Thing
HostSimple	Host	HumanAgent	Agent
HostCoupling	Host	ArtificialAgent	Agent
VirtualizedResource	Thing	VO	Thing
X509	Thing	CognitiveEnv	Thing
Service	Thing	Capacity	Thing

Using an ontology to describe GRID-MAS integrated models is interesting because we can describe using the same formalization both the model and its instances. Designers of GRID-MAS integrated models may instantiate AGIO concepts in order to formalize their systems and check their consistency thanks to AGIO rules, as it is illustrated in Sect. 4.

One pre-requisite to any collaboration is to have mutual understanding about things the collaboration is dealing with. The intrinsic elements agents need a mutual understanding about, are themselves and the world in which they evolve. It is called a meta-description. Actually, AGIO is a language for expressing such a meta-description because agents of GRID-MAS integrated systems can use it in order to represent the world in which they exist and exchange services. AGIO allows agents to agree on what they are, what is a service, a host, etc.

Table 2. AGIO' relations and types

RELATION	TYPE	DOMAIN	RANGE
<i>couples</i>	relation	HostCoupling	Host
<i>virtualizes</i>	relation	VirtualizedResource	Host
<i>reifies</i>	relation	ServiceContainer	VirtualizedResource
<i>holds</i>	relation	Agent \cup Host	X509
<i>belongs</i>	relation	Agent	VO
<i>interacts</i>	function	Agent	Agent
<i>uses</i>	relation	Agent	Service
<i>provides</i>	function	Agent	Service
<i>exchanges</i>	relation	Agent	Agent
<i>instantiates</i>	relation	NormalService CognitiveEnv	Service CognitiveEnv
<i>includes</i>	function	ServiceContainer	CAS \cup NormalService
<i>handles</i>	relation	VO CAS	CAS CAS \cup NormalService
<i>interfaces</i>	function	Service	Capability
<i>executes</i>	function	CognitiveEnv	Capability
<i>owns</i>	function	ArtificialAgent	CognitiveEnv

3.3 OWL/Protégé Implementation

AGIO was implemented in Protégé. For instance, the class⁴ *Agent* and the property *interfaces* are defined in OWL as:

```
<owl:Class rdf:about="#Agent">
  <owl:disjointWith rdf:resource="#X509" />
  ...
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="belongs" />
      </owl:onProperty>
      <owl:someValuesFrom>
        <owl:Class rdf:about="#VO" />
      </owl:someValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  ...
  <owl:ObjectProperty rdf:about="#interfaces">
    <owl:inverseOf rdf:resource="#isInterfacedBy" />
    <rdfs:range rdf:resource="#Capability" />
    <rdfs:domain rdf:resource="#Service" />
  </owl:ObjectProperty>
```

An example of integration rule may be: *an agent must hold a X509 certificate in order to be member of a VO*. It is expressed in SWRL as:

$$Agent(?x) \wedge X509(?y) \wedge VO(?z) \wedge \mathbf{belongs} (?x,?z) \rightarrow holds(?x,?y)$$

4 Example of Scenario with AGIO

Data mining scenario. [14] proposes a Grid based architecture for supporting knowledge discovery processes such as data mining, data analysis, etc. The *Knowledge Grid* is a set of layers upon the GRID, allowing parallel and distributed knowledge discovery. This knowledge Grid architecture is based on a set of services constructed on the top of basic core GRID services. In this section, we present a "data mining" scenario which considers an artificial agent that provides users with a data mining Knowledge Grid service. This is a good example of a potential GRID-MAS integrated system, because Knowledge Grid services are typically the kind of services that could be very enhanced by intelligent agent abilities.

AGIO's instances of the scenario. Let us consider the following elements: BOB, LUCY and DMAGENT are three instances of *Agent*. They *belong* VO1, an instance of VO (Virtual Organization). They *hold* respectively XB, XL, and XD, three instances of X509. DMAGENT *provides* DMSFACT, DMSB and DMSL, three instances of

⁴ Protégé calls class and property respectively concept and relation.

NormalService. They are *interfaced* by respectively DMGENERICCAP, DMCB and DMCL, and they are *executed* in CEGLOBAL, CELOCALB and CELOCALL (Cognitive Environment), all three *owned* by DMAGENT. DMSFACT *instantiates* DMSB and DMSL and CEGLOBAL *instantiates* CELOCALB and CELOCALL. CAS1, an instance of CAS, *handles* itself, DMSFACT, DMSB, DMSL and is *handled* by VO1. All these services are *included* in SC1, an instance of ServiceContainer. The rest of the instances (VirtualizedResource, Host, etc.) are not specified here. In particular, the VirtualizedResource available for DMSFACT, DMSB and DMSL may come from a set of different StorageResource that DMAGENT integrates and process via its data-mining services.

BOB wants to analyze some elements in the common shared set of data proposed by GRID. LUCY looks for a specific data pattern in same set of data. They both have different objectives. Thus, BOB and LUCY respectively *uses* DMSB and DMSL. These instances are customized, thank to DMAGENT intelligent and interactive abilities. The two different interfaced capabilities are executed in two different conversation contexts dedicated for their users. These Knowledge Grid services are significantly enhanced by agent abilities. [6,3] detail the ways an agent can dynamically generate services to another one, following a conversation.

Consistency. On a given scenario such as the "data-mining" scenario, one can check the consistency of the set of instances thanks to the SWRL rules implemented in AGIO e.g., if exists the right set of instances for a given property. Each conjunction may be considered as a predicate. For example:

$$\begin{aligned}
 & Agent(DMAGENT) \wedge CAS(CAS1) \wedge Service(DMSB) \wedge \\
 & VO(VO1) \wedge CognitiveEnv(CELocalB) \wedge Capability(DMcB) \wedge \\
 & \quad \mathbf{provides(DMAGENT, DMSB)} \rightarrow \\
 & owns(DMAGENT, CEMLocalB) \wedge executes(CELocalB, DMcB) \wedge \\
 & \quad interfaces(DMSB, DMcB) \wedge belongs(DMAGENT, VO1) \wedge \\
 & \quad \quad handles(VO1, CAS1) \wedge handles(CAS1, DMSB)
 \end{aligned}$$

5 Related Work and Discussion

There is an increasing amount of research activity in GRID and MAS convergence taking place.⁵ The Control of Agent-Based Systems (CoABS) project [15], proposed in 1999 by the DARPA, is the first research initiative in GRID-MAS integration. In this project, priority was given to GRID development, but the participants already envisage a combination of GRID and MAS domains. The use of agents for GRID was very early suggested in [16]. The authors specifically detail how agents can provide a useful abstraction at the Computational Grid layer. MAS has also been established as a key element of the Semantic Grid [17]. More recently, the why GRID and MAS need each other as been established by [1].

⁵ See, for example, *Agent-Based Cluster and Grid Computing* workshops, *Smart Grid Technologies* workshops, the *Multi-Agent and Grid Systems* international journal.

Using MAS principles to improve core GRID functionalities represent the main part of related work. For example, MAS-based GRID approaches for resource management use agents for an effective management of the vast amount of resources that are made available within a GRID environment as they have, for example, excellent trading and negotiation abilities (negotiation between resource agents and allocator agent) e.g., [15,18,19]. Moreover MAS can use the Grid Laboratory Uniform Environment (GLUE Schema⁶) to access easily to resources. The GLUE Schema is a description and a representation of GRID heterogeneous resources. Agents, using this GLUE Schema, are able to retrieve resources without taking into account the resources origin. It would be a good way to share the status and availability of resources and resources themselves.

Another example is the use of MAS for VO management i.e., formation, operation and dissolution of VOs. The main work in this domain is the Grid-enabled Constraint-Oriented Negotiation in an Open Information Services Environment (CONOISE-G) project [20].

Some work has also been proposed in using agents to enhance Web/Grid services or integrating the two approaches. In particular, in order to connect MAS communication approaches with business process management approaches e.g., [21,22,23,4]

However, none of these works propose a real integration of MAS and GRID. Rather, they focus on how MAS and AI techniques may enhance core GRID functionalities. Our vision of a GRID-MAS integration is not a simple interoperation of the technologies (it goes beyond a simple use of one technology to enhance the other). Besides, describing the integration of GRID and MAS by means of an ontology sets a new formal foundations to the integrating ideas.

AGIO ontology focuses on modelling an integration of GRID-MAS. We can use the OntoGrid⁷ to formalize more precisely GRID concept, but the aim of our work is to present an integration of GRID-MAS using general concepts of GRID. We do not have the ambition to represent all concepts linked to GRID, but to succinctly express the way that MAS and GRID are integrated each other.

The advantages of the GRID-MAS integrated model formalized by AGIO are precisely detailed in [3]. We briefly describe some of them here. There is no real standard in the MAS community to describe agent capabilities. Interfacing them as Grid services is thus a potential step towards standardization. This integrated model does not restrict MAS or GRID in any way. Everything feasible with MAS or GRID today still holds. VO management benefits from both GRID and MAS organizational structure formalism, e.g., Agent-Group-Role [7], CAS service, X509 certificate, etc. Service exchange benefits from the important agent communication abilities, e.g., dealing with semantics, ability to have a conversation, etc. The challenge of modelling dynamic agent conversations becomes the same as the one of dynamically composing and choreographing services in business processes. The model subsumes a significant number of the MAS-based GRID approaches mentioned before thanks to the reflexivity of GRID, which defines some GRID core functionalities as (meta-)Grid services (e.g., service container, CAS). Therefore, GRID and MAS would appreciate a common ontology which:

⁶ <http://glueschema.forge.cnaf.infn.it/>

⁷ <http://www.ontogrid.net>

- describes simply and clearly key concepts and their integration;⁸
- uses the same terms and representations for an identical concept e.g., VO and group, choreography of service and agent conversation, role and service.
- rigorously fixes the integration rule;
- may help researchers of GRID and MAS communities to specify and model their GRID-MAS integrated applications and systems by instantiating AGIO concepts (pass from a procedural ontology to a descriptive one);
- would promulgate the development of GRID-MAS integrated systems by proposing a uniform way of describing GRID and MAS together.

In a sake of simplicity, the paper presents only some part AGIO's rules or OWL elements. We invite the reader to refer to [24], for a complete specification of AGIO, or to [3] for a complete specification of AGIL.

6 Conclusion and Perspectives

Even if using agents for GRID was very early suggested [15,16,17], Foster et al. [1] propose the real first step in GRID-MAS integration as it examines work in these two domains, firstly to communicate to each community what has been done by the other, and secondly to identify opportunities for cross fertilization as they explained how GRID and MAS developed significant complementarities. The work proposed by [2], [3] and this paper suggests a second step by proposing a GRID-MAS integrated model. In particular, in this paper, we formalize this model by means of an ontology. AGIO describes concepts of the GRID-MAS integrated model, as well as relations between each of these concepts. Moreover, relations are strengthened by SWRL rules which guarantee coherence.

Some perspectives for the future of AGIO may be:

- To add new concepts, properties and rules according of the evolution of the GRID-MAS integrated model e.g., different types of X509 certificates (proxy's, certification authority's);
- To connect AGIO with other meta-ontologies used to define service, GRID or MAS concepts;
- To experiment the ontology with large-scale GRID-MAS integrated systems;
- To integrate in AGIO other different agent representation approaches;
- To classify services with their intrinsic properties (e.g., statefulness, transient, multipoint);
- To define tools that assist designers by automatically checking, during system development, the consistency of the ontology.

References

1. Foster, I., Jennings, N.R., Kesselman, C.: Brain meets brawn: why Grid and agents need each other. In: 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS'04. Volume 1., New York, NY, USA (2004) 8–15

⁸ In particular, GRID needs a description language that summarizes and rigorously explains GRID concepts. Without considering the agent side, AGIO may play also this role.

2. Jonquet, C., Dugenie, P., Cerri, S.A.: Service-based integration of Grid and multi-agent systems models. Research report 06012, University Montpellier II, France (2006) www.lirmm.fr/~jonquet/Publications.
3. Jonquet, C., Dugenie, P., Cerri, S.A.: AGIL specifications. Research report 06030, University Montpellier II, France (2006) www.lirmm.fr/~jonquet/Publications.
4. Singh, M.P., Huhns, M.N.: Service-Oriented Computing, Semantics, processes, agents. John Wiley & Sons (2005)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The physiology of the Grid: an Open Grid Services Architecture for distributed systems integration. In: Open Grid Service Infrastructure WG, Global Grid Forum, The Globus Alliance (2002)
6. Jonquet, C., Cerri, S.A.: The STROBE model: Dynamic Service Generation on the Grid. Applied Artificial Intelligence, Special issue on Learning Grid Services **19**(9-10) (2005) 967–1013
7. Ferber, J., Gutknecht, O., Michel, F.: From agents to organizations: an organizational view of multi-agent systems. In Giorgini, P., Müller, J.P., Odell, J., eds.: 4th International Workshop on Agent-Oriented Software Engineering, AOSE'03. Volume 2935 of Lecture Notes in Computer Science. Springer-Verlag, Melbourne, Australia (2003) 214–230
8. Ferber, J.: Multi-agent systems: an introduction to distributed artificial intelligence. Addison Wesley Longman, Harlow, UK (1999)
9. Wooldridge, M.: An introduction to multiagent systems. John Wiley & Sons, Chichester, UK (2002)
10. Comito, C., Talia, D., Trunfio, P.: Grid services: principles, implementations and use. Web and Grid Services **1**(1) (2005) 48–68
11. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition **5**(2) (1993) 199–220
12. : OWL Web Ontology Language use cases and requirements. W3c recommendation, World Wide Web Consortium (2004)
13. O'Connor, M., Knublauch, H., Tu, S.W., Grosz, B.N., Dean, M., Grosso, W.E., Musen, M.A.: Supporting Rule System Interoperability on the Semantic Web with SWRL. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: 4th International Semantic Web Conference, ISWC'05. Volume 3729 of Lecture Note in Computer Science., Galway, Ireland, Springer-Verlag (2005) 974–986
14. Cannataro, M., Talia, D.: The Knowledge Grid. Communications of the ACM **46**(1) (2003) 89–93
15. Manola, F., Thompson, C.: Characterizing the agent Grid. Technical report 990623, Object Services and Consulting, Inc. (1999)
16. Rana, O.F., Moreau, L.: Issues in building agent based computational Grids. In: 3rd Workshop of the UK Special Interest Group on Multi-Agent Systems, UKMAS'00, Oxford, UK (2000)
17. Roure, D.D., Jennings, N., Shadbolt, N.: Research agenda for the Semantic Grid: a future e-science infrastructure. Technical report, University of Southampton, UK (2001) Report commissioned for EPSRC/DTI Core e-Science Programme.
18. Martino, B.D., Rana, O.F.: Grid performance and resource management using mobile agents. In Getov, V., Gerndt, M., Hoisie, A., Malony, A., Miller, B., eds.: Performance Analysis and Grid Computing. Kluwer (2003) 251–263
19. Cao, J., Spooner, D.P., Jarvis, S.A., Nudd, G.R.: Grid load balancing using intelligent agents. Future Generation Computer Systems **21**(1) (2005) 135–149
20. Patel, J., Teacy, W.T.L., Jennings, N.R., Luck, M., Chalmers, S., Oren, N., Norman, T.J., Preece, A., Gray, P.M.D., Shercliff, G., Stockreisser, P.J., Shao, J., Gray, W.A., Fiddian, N.J., Thompson, S.: Agent-based virtual organisations for the Grid. Multiagent and Grid Systems **1**(4) (2005) 237–249

21. Buhler, P.A., Vidal, J.M., Verhagen, H.: Adaptive workflow = Web services + agents. In: International Conference on Web Services, ICWS'03, Las Vegas, NV, USA, CSREA Press (2003) 131–137
22. Ardissono, L., Goy, A., Petrone, G.: Enabling conversations with Web services. In: 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems, AAMAS'03, Melbourne, Australia, ACM Press (2003) 819–826
23. Greenwood, D., Calisti, M.: Engineering Web service - agent integration. In: IEEE Systems, Cybernetics and Man Conference, The Hague, Netherlands, IEEE Computer Society (2004)
24. Duvert, F.: An ontology of Grid and Multi-agent systems integration. Master's thesis, University Montpellier 2, Montpellier, France (2006)

COMINF 2006 PC Co-chairs' Message

Community informatics, also known as community networking, electronic community networking, community-based technologies or community technology, refers to an emerging set of principles and practices concerned with the use of information and communications technologies (ICTs) for personal, social, cultural or economic development within communities; for enabling the achievement of collaboratively determined community goals; and for invigorating and empowering communities in relation to their larger social, economic, cultural and political environments.

From an academic and research perspective, community informatics can be seen as a field of practice in applied ICT. It brings together the practices of community (and economic and social) development with insights from fields such as sociology, planning, development studies, women's studies, library and information sciences, management information systems, and management and computer sciences. Its outcomes – community networks and community-based ICT-enabled service applications – are of increasing interest to grassroots organizations, NGOs and civil society, governments and multilateral agencies, and the private sector, among others. Self-organized community ICT initiatives spanning the range of application areas including health, social and economic development, small business, environmental management, and local governance have been emerging world-wide with the objective of harnessing ICTs for developing social capital, poverty alleviation and empowerment at the local level. In addition, collaborative communities enabled with ICTs are helping to bridge organizational boundaries, ensuring more effective and efficient forms of collaboration in and between stakeholders from business, government, education, and civil society.

Most recently ICTs are playing a key role in enabling many types of virtual or hybrid communities. The resulting socio-technical systems, however, are very complex and continuously evolving. The intricate interactions between community requirements and their enabling technologies are still ill understood. In particular, there is a huge gap between those who understand the complexities and dynamics of communities and those information technologists who can build the technologies and systems that can catalyze and enable communities into more effective action.

In this community informatics workshop, our intention was to gather both researchers and practitioners interested in the modeling and analysis of community requirements, the design and implementation of community-based ICTs and community information systems, and the evaluation of these technologies, as a way of contributing to the effective use of information systems at the community level. Some examples of topics of interest which were discussed in this workshop include: community requirements modeling and analysis; technologies for enabling communities (weblogs, discussion fora, portals, nology) and locally significant broadband applications

In all, 33 papers were received, 16 of which were accepted for presentation at the workshop and inclusion in the proceedings. Each paper underwent either 2 or 3 peer reviews. We would like to thank the authors and members of the Program Committee for the quality of their contributions and their commitment to making this workshop a success.

The time has come to ground and interweave the emerging discipline of community informatics into the mainstream of computer and information science. Recognizing the significance of communities and particularly community-based applications as one of the quality of the papers submitted for the workshop. Our plan is to make this workshop a recurring event. By examining the link between information technology and community, we hope to make a broad contribution towards more developed theory and a more focused application of the powerful potential of ICTs in response to the range of societal and grassroots contexts and socially meaningful issues.

August 2006

Aldo de Moor
Vrije Universiteit Brussel, Belgium
Michael Gurstein
Community Informatics Research Network, Canada

Community Informatics and Human Development

William McIver Jr.

National Research Council Institute for Information Technology,
46 Dineen Drive, Fredericton, New Brunswick, Canada E3B 9W4

Abstract. A global crisis in human development and prosperity exists. Its constituents include extreme poverty, illiteracy, gender imbalances, and armed conflict. One response has been the creation of the Millennium Development Goals. The recent United Nations World Summit on the Information Society affirmed support for linking the further development of the global information society to these goals. This paper argues that community informatics has a specific role to play in contributing to the realization of the Millennium Development goals and the improvement of human development in general.

Keywords: community informatics, human development, poverty, Millennium Development Goals.

1 Our World

Social, scientific, and political developments over this last millennium have brought remarkable progress to the overall human condition. Average life span, literacy, agricultural production, and the ability to manage large social organizations have all improved greatly in the aggregate. However, it is difficult to say that our world is not in crisis, whether viewed from the perspective of sustained and recurring historical processes or as a conjunction current social, material, and environmental conditions.

We are in an epoch where significant inequalities remain and poverty, disease, and military conflict are devastating communities at significant levels. Over 20% of people in the world live in extreme poverty [1]. Over 100 million children are not receiving basic schooling [1]. Child mortality rates in some parts of the world are as high as 29 times that of most developed countries [1]. Over 1 million people die annually because of lack of access to clean water [1, 2, 3]. By the year 2015, HIV/AIDS is expected to have reduced populations in developing countries by 100 million. In many of these countries whole public service sectors have already been devastated by the epidemic, thereby exacerbating existing social problems [1]. Women and girls are at heightened risk in most human development indicators, including health and access to education. More than 20 major wars and other armed conflicts have occurred over the past decade, causing sustained damage to social and political infrastructure. Genocide has been formally recognized in several of these wars (see [6] and [7]). More details are given in an appendix.

Linkages between all of these issues reveal added complexities. For example, women who are less educated are more likely to become infected with HIV, while access to primary education for many children is being reduced significantly as teachers and other professionals die from AIDS.

Responding to the troubling status of human development, member states of the United Nations agreed unanimously to the Millennium Declaration which set goals for addressing key human development indicators to be achieved by the year 2015 [8]:

- Goal 1.* eradicate extreme poverty and hunger;
- Goal 2.* achieve universal primary education;
- Goal 3.* promote gender equality and empower women;
- Goal 4.* reduce child mortality;
- Goal 5.* improve maternal health;
- Goal 6.* combat HIV/AIDS, malaria, and other diseases;
- Goal 7.* ensure environmental sustainability; and
- Goal 8.* develop a global partnership for development.

These are now widely known as the Millennium Development Goals (MDGs). They have become the primary metrics for measuring human development. They have also become a major component of agreements in international forums, including the recent United Nations World Summit on the Information Society (WSIS). For many people, the MDGs are the current benchmarks against which policy and action in many domains should be judged. Governments and civil society in the WSIS, for example, decided that MDGs should be the basis of its plan of action: “Our challenge is to harness the potential of information and communication technology to promote the development goals of the Millennium Declaration” [9].

It is difficult to dissent from the claim that our world is in crisis if the current status of the MDGs is examined. The United Nations Development Programme reported in its 2005 Human Development Report that at 10 years from the 2015 Millennium Development Goals, “if current trends continue, the MDGs will be missed by a wide margin what was possible” [1].

Technology can make positive contributions in the attempt to meet these goals. Stephen Lewis, former United Nations Ambassador and Special UN Envoy on HIV/AIDS in Africa, when asked about positive applications of ICTs in responding to HIV/AIDS in the developing world cited several examples: applications of basic mobile telephone technology in a number of countries, including Botswana; the use of computers to track medication and patient records in Swaziland; and the implementation of distance learning by the Royal College of Nursing in the UK to carry out remote training of medical staff [10]. Experience has taught the computing professions that if examples such as these are successes, they are as much, if not more, the result of careful social processes as they are the result of the unique characteristics of the particular technologies being used. The traditional field of management information systems and the newer field of community informatics provide different perspectives on the social processes necessary for conducting technology design and implementation processes. This paper attempts to make the case that community informatics has a specific role to play in facilitating technological contributions to the improvement of human development.

Community informatics is an emerging discipline that is developing ICT analysis, design and deployment techniques that are geared toward the unique requirements of communities. This is in contrast to the field of management information systems, which was developed for organizational contexts. The linkages between community informatics and human development can take several trajectories. One trajectory relates the fundamental importance of information in addressing any societal issue to

the inherent potentials in information and communication technologies (ICTs) for facilitating the exchange and processing of information. It is held as axiomatic that communication is a fundamental social process and a foundation of all social organization. Thus, ICTs must not be seen only at a technical level for their processing capabilities; they must be seen at a social level as enablers – for good or bad -- of social processes and social organization. In particular, ICTs are potential enablers of effective formation and communication of information that is vital to addressing human development issues. A second trajectory is a corollary of the first, which is that if ICTs are to be viewed as integral to the search for solutions to problems of human development, then community informatics is among the key disciplinary approaches that are necessary for solution-seeking. Traditional management information systems approaches are designed for organizations and are not necessarily well-suited to addressing the unique requirements of communities.

This paper does not seek to be a comprehensive treatise on community informatics nor to focus on specific MDGs and case studies. It represents the beginning of a process for the author. This paper has two goals: (1) to motivate the development of a more comprehensive vision and a plan for contributing community informatics expertise to the improvement of human development and (2) to propose an initial set of tasks and challenges for accomplishing the first goal. Sections 2 and 3 address the first goal. Section 4 addresses the second goal.

2 Community Informatics as Praxis

What difference can community informatics practitioners make in all of this? The author still encounters arguments from peers and colleagues suggesting that computing and information sciences have no relationships nor obligations to social and political spheres. Perhaps one indication of this type of thinking is the fact that the two major academic computing societies in the world, the Association for Computing Machinery and the IEEE Computer Society, had little or no official representation at the recent United Nations World Summit on the Information Society. In contrast, non-governmental organizations (NGOs) from a wide variety of interests participated vigorously in the policy debates there.

This separation between academic and practical endeavors is understandable on the grounds of disciplinary focus, but the bridging function that science provides between policymaking and implementation – to both positive and negative effect -- is not new. The problem is that research, engineering, and implementation processes produce the ICTs which give constant rise to serious policy issues. The response has usually been to address only the direct and indirect effects on society of ICT production and use, and, to a lesser extent, how the characteristics of certain ICTs might be leveraged to either address human needs or to violate human rights. Thus, the processes and professions that produce ICTs might be said to exist as “black boxes” relative to the social impacts they make. In an abstracted view of this, the computing and engineering professions are arguably the most responsible for generating the technological artifacts and complex systems that yield the transitive social, political, and economic phenomena to which society must later respond. Yet, these professions and their approaches have been subject to only minor scrutiny and calls for accountability.

This call to link community informatics research and praxis more closely to global society's most serious problems is not a call to technological constructionism. Most people within technology and policy sectors -- and society at-large -- now recognize that technology is not infallible and that technology cannot and should not be counted on alone to solve social and political problems. It can be argued, however, that given the phenomenal evolution of technologies for enhancing the production and communication of knowledge, the computing professions have an obligation to examine what contributions they can make to addressing major social issues. What is clear to practitioners who have gravitated to the disciplines of social and community informatics is that ICTs have become inextricably linked with almost all facets of human society and they have the potential to bring both benefit and harm to it.

This calls into question the nature of community informatics. The author has previously defined community informatics as [11]:

... an interdisciplinary field concerned with the development, deployment and management of information systems designed with and by communities to solve their own problems. It is arguably a part of social informatics, which has been defined by Kling [12] as 'the interdisciplinary study of the design, uses and consequences of information technologies that takes into account their interaction with institutional and cultural contexts.'

Community informatics can be seen in the context of social informatics as a disciplinary site focusing specifically on the roles and relationships between ICT and communal environments and their unique requirements.

This definition is not enough, however. Community informatics must be seen clearly to inform and encourage a praxis. That is, teaching and research within the discipline must be directed more vigorously to the practical application of its knowledge. Such a praxis should be focused, in particular, on major societal issues such as those represented by the MDGs. Community informatics is uniquely suited among the computing and information science disciplines in this context given its specific attention to social, economic, cultural and other facets of communities in understanding the ways that ICTs should and should not be used.

It may still be difficult for some people to contemplate actionable linkages between science and engineering and issues such as the MDGs, but this is not a new type of framework for science and engineering. The organization Doctors Without Borders / Médecins Sans Frontières (MSF), founded in 1971, is an important example of a professional community that has resolved the linkages between its general scientific and ethical obligations and specific critical social needs -- in this case the delivery of medical care in emergency situations [13]. Since 2002, there is now also Engineers without Borders -- International, an organization that has explicitly sought to contribute engineering expertise toward meeting the MDGs [14]. Analogous efforts should be contemplated within the computing professions.

Community informatics is itself an acknowledgment of more than the non-value-neutral and fallible nature of ICT. It is a recognition that the idea of purely technical solutions to societal problems is a fallacy and, further, that the seeking of technical solutions must necessarily be a social process. In this context, advanced ICTs such as the Internet must be considered only "candidate solutions," in the parlance of systems analysis, within a community informatics approach. A better phrase might be

“candidate contribution.” That is, advanced ICTs must not automatically be assumed to provide the best solutions. Instead, ICTs in general must be viewed only as potential components of overall solutions to a given problem, where other non-technical and social components are assumed to play the leading roles. Thus, the process of finding candidate contributions is a holistic one. This process includes, but is not limited to, the following steps: (1) understanding the community context to which ICTs might be applied; (2) developing appropriate and sustainable models for the socio-technical systems into which such ICTs are to be integrated; and (3) selecting and appropriating ICTs based on knowledge gained from parts (1) and (2) of this process. The overriding concern then is to select technologies that are suitable and appropriate to a community given social, cultural, sustainability and economic factors. To achieve this, it is necessary to be open to the full range of communication modalities and technologies, including simpler and older technologies.

The pressing nature of the problems outlined in the first section calls for every tractable and useful strategy to be brought to bear to solve them, including the application of ICTs. Taken from a traditional management information systems (MIS) perspective, where the materials and expert computing professionals required to develop systems are often assumed to be abundant, the identification of feasible technical contributions to meeting goals such as those in the Millennium Declaration might be seen as relatively straightforward. The realities for developing countries are vastly different. ICT resources and computing professionals are usually both scarce in the developing world. The characteristics of communities in general are highly unique relative to organizations and, therefore, the development of ICTs for communities warrants a special focus [15, 16]. Historically, MIS project failures exceed successes by a significant margin, thus, special care must be taken in systems design processes within community contexts since the economic, political, and social tolerances for failure are usually not as high in communities as they are in large industrial and governmental organizations [17].

3 Our Technology

In contrast with the poverty and inequality that exists today, there is a growing abundance of information technology and services in the developed world. This is a technological epoch where the usage, reach and capabilities of ICTs continue to accelerate at a remarkable rate. It cannot be argued, therefore, that there not sufficient resources to try to use ICTs to contribute solutions to these problems.

Dispersion of ICTs is but one indicator of this progress. According to the International Telecommunication Union [18] the period between 1994 and 2004 saw in the developing world alone increases in the proportion of¹:

- fixed telephone lines per inhabitant, from 4.4 to 12.8 per 100 inhabitants;
- mobile telephone subscribers, from 0.19 to 18.8 per 100 inhabitants; and,
- Internet users, from 0.03 to 6.7 per 100 inhabitants.

Correspondingly, in the developed world between 1994 and 2004 there were, according to the ITU [18], increases in the proportion of:

¹ These ITU data were collected from 206 economic areas with populations greater than 40,000 (ITU May, p.9).

- fixed telephones lines from 48.8 to 53.5 per 100 inhabitants;
- mobile telephone subscribers from 5.20 to 76.8 per 100 inhabitants; and
- Internet users from 2.18 to 53.8 per 100 inhabitants.

Nielsen/NetRatings reported that the total numbers of online searches increased 55% over the 12 months starting in December 2004 [19].

Innovation is perhaps a more important indicator of the potential of ICT research to address at least some of the major human development problems. Patents are but one indicator of innovation. Phenomenal increases in ICT patent filings have been seen over the past decade. Information and communication technology (ICT) patents filed by the 30 member countries of the Organisation for Economic Co-operation and Development (OECD) with the European Patent Office (EPO) made up 16.4% of their total national patents [20]. By the year 2000, ICT patent filings with the EPO made up over 34% of their total national patents [21]. Between 1990 and 1995, investment in ICT contributed an average of 34% to the growth of GDP across member countries. This increased to 55% between 1995 and 2003 [22].²

Even more phenomenal innovation can be seen outside of the restrictive, and arguably counter-productive, application of patent and copyright regimes. The community of free and open source software (FOSS) has seen phenomenal growth over the past decade. FOSS has been responsible for making numerous free (or low total cost of ownership) solutions to individuals and communities, including operating systems, document processing applications, and web-based content management systems. Open specification hardware, in analogue to FOSS, is perhaps less familiar to many, but is hardly new. The seeking of ICT-based solutions has had to include consideration of low cost, public domain or open source solutions. Open technologies will enable communities to be more self-sufficient in replicating, maintaining and enhancing ICT-based development projects. This approach allows a community to have complete access to the internal workings of the technologies they use. Many open technologies are have no cost licensing, alleviating developers of much of the costs of acquiring technologies.

The crucial point is that the costs, processing power, portability, energy consumption rates of commodity ICTs are now such that there is little excuse for not vigorously searching for ways that they can contribute to addressing major human development issues.

4 Tasks and Challenges

Disciplines such as computer science and mathematics periodically develop “grand challenge” problems, sets of problems whose solutions are necessary to make major progress in a field and which possibly offer applications that would significantly advance society. One example is supercomputing [23]. The challenge in developing technological contributions to solving human development issues is unique relative to traditional management information systems in that they must be socially and culturally appropriate and operationally, economically, environmentally, and socially

² Only 19 of 30 member countries are accounted for in the data cited here due to a reporting threshold set by the OECD.

sustainable. This is especially true for developing and least developed countries, where resources and training are certainly scarcer than in most communities.

A significant amount of community informatics-related praxis targeting issues in the MDG sphere has already taken place. This can be seen in numerous FOSS projects, the development of telecentres around the world, and numerous ICT training and policy projects. An examination of projects developed by NGOs that participated in the WSIS provides a good sample of this type of work.

Still, there is much to be gained from developing a systematic community informatics approach wherein best practices from “work on the ground” can be shared and taught within established research and pedagogical networks. Two major tasks must be pursued for such a process to commence: systematic stocktaking and the development of a comprehensive community informatics curriculum.

Stocktaking is a necessary part of any institutional programme, in this case the academic discipline of community informatics. An ongoing process is needed to assess its contributions, shortcomings, and strategic directions. Issues in the MDG sphere should be a special focus of this stocktaking.

The development of a community informatics curriculum has been discussed within the Community Informatics Research Network (CIRN), arguably the main body for this discipline. This work must be completed. Such a curriculum is a necessary framework for stocktaking to be performed and for establishing strategic directions for the discipline. An analogous relationship can be seen in the Association of Computing Machinery's computing curricula, which have for over forty years been critical components in the creation of institutional road maps for computer science and related areas of engineering [24].

Several research and development challenges exist for which efforts must be redoubled -- some of these are already well-known: furthering the understanding of sustainability, development of educational resources for developing and least developed countries, and addressing rural and remote issues.

The development of sustainable ICT contributions to MDG sphere issues must address constraints in several dimensions. These include, but are not limited to: economic sustainability, environmental sustainability, and the development of productive forces within target communities. Economic sustainability of ICT projects is crucial and has been addressed extensively. Greater attention must be given to creating ICT life cycles which are environmentally sustainable. Significant progress has been made over the past two decades in reducing energy consumption, including processors, energy-aware software behaviour, and applications of LED technology. Practical, environmentally-sustainable solutions to power generation, particularly for remote applications are now more widely available. One interesting example is the recent development of an integrated solar powered WiFi solution designed specifically for constraints faced by developing countries [25]. There also exists an emerging telecommunication sector that is focused on developing low cost wireless communications devices, such as VoIP mobile telephones; and systems that permit the implementation of community-based telephone and cable television cooperatives [26, 27].

Work remains, however, on managing the end of the life cycle for hardware. Many developing and least developed countries have themselves become victim to the end of the supply chain of hardware consumption through dumping and the harvesting of parts under dangerous labor conditions. What is required is a commitment to evolving

“cradle to cradle” design of hardware and system life cycles, where these activities are no longer environmentally damaging nor hazardous to workers [28].

Sustainability within community informatics must also include development of productive forces within target communities themselves for creating and managing ICT. Participation is a necessary but insufficient condition for ICT design to be effective. Among the greatest threats of new technologies is that they have the potential to perpetuate and expand existing power relations and inequalities, as well as to enable new forms of state repression. To empower communities to respond to and avoid these threats, community informatics must enable a fully democratic and consensual process. That is, it must facilitate more than political democracy embodied in participatory ICT design approaches. Community informatics must go beyond this to enable people to share control of the decision making around the economic, cultural, environmental and other issues regarding ICT-based projects. More fundamentally, community informatics must empower communities that contemplate ICT-based solutions to develop their own productive forces within the information society so that they can control the modes of production that evolve within it and, thereby, have the possibility of preventing and responding to its threats.³ FOSS as a mode of production is a prime example. In this case, it can enable communities to develop their own means of creating software. A current example of these principles being put into practice is the decision by the Extremadura region of Spain to implement ICT solutions that are completely open source, including the use of the Open Document Format (ODF) [29].

The fundamental requirement for creating this type of productive capacity within developing and least developed countries is the creation of educational resources. Text books are too expensive for many students in developed countries; access to such material in developing and least developed countries is all but impossible. The creation of free educational literature is the only practical approach to solving this problem. One new model is the Global Text Project, which plans to write electronic text books across all major academic disciplines using collaborative web technologies [30]. Community informatics can make significant contributions to such efforts not only in developing content, but in supporting and improving these types of educational models through the development of low cost ICTs that are sustainable in developing countries, including rural and remote regions. Specific research problems, include improving support for natural language translation of texts and optimizing document caching and delivery mechanisms to operate in low bandwidth and low connectivity environments.

5 Conclusions

This paper is the beginning of a process. It is a call for a consensual and democratic process to commence within existing networks, such as CIRN, to more fully develop this vision of linking community informatics more closely to the pursuit of the Millennium Development Goals and the improvement of human development in general. The community informatics community has a unique opportunity following the WSIS and the processes it has set in motion to help shape information societies for

³ The author has long been inspired by various writings of Walter Rodney [31] and Amílcar Cabral [32] in this context. For comprehensive treatment of these issues see Powell [33].

human needs. It can, in particular, help to focus ICT research, design, and implementation on addressing the most pressing human development issues. This includes directing parts of its research and pedagogy on developing a more effective praxis for this purpose and on continuing to evolve its networks of practitioners. Finally, the community informatics community must also oppose designs and applications of ICTs that perpetuate social, economic, and political inequality and contribute to armed conflict.

Acknowledgments. The author thanks Aldo de Moor and Michael Gurstein for their in-depth comments on this paper.

References

1. United Nations Development Programme (UNDP). (2005). *Human Development Report 2005: International cooperation at a crossroads: Aid, trade and security in an unequal world*, UNDP, 2005, pp. 17, 28, 34, 44, 45; <http://hdr.undp.org/reports/global/2005>.
2. World Water Council, *Water Crisis*, March 2, 2006; <http://www.worldwatercouncil.org/index.php?id=25>.
3. World Health Organization (WHO), *Water, sanitation and hygiene links to health: Facts and figures updated November 2004*, November 2004; http://www.who.int/water_sanitation_health/publications/facts2004/en/.
4. United Nations Programme on HIV/AIDS (UNAIDS), *2006 Report on the global AIDS epidemic*, UNAID, S/06.20E (English original, May 2006), pp. 13, 81.
5. Lewis, S., *Race Against Time*, Toronto: House of Anansi Press, 2005.
6. Human Rights Watch, <http://www.hrw.org>.
7. U.S. Central Intelligence Agency, Field Listing – Background; <https://www.cia.gov/cia/publications/factbook/fields/2028.html>.
8. United Nations, *United Nations Millennium Declaration*. Draft Resolution referred by the General Assembly at its fifty- fourth session, Item 61(b) of the provisional agenda, 2000; <http://www.un.org>.
9. United Nations. "Declaration of Principles: Building the Information Society: a Global Challenge in the New Millennium." World Summit of the Information Society. Document WSIS-03/GENEVA/DOC/4-E, 12 December 2003 Original: English, p. 1.
10. Lewis, S., *Personal communication*, Fredericton, New Brunswick, Canada, April 28, 2006.
11. McIver, Jr., W. J., "A Community Informatics for the Information Society," IN Seán O'Siochru and Bruce Girard (eds.): *Communicating in the Information Society*. Geneva, Switzerland: United Nations Research Institute for Social Development (UNRISD), 2003.
12. Kling, R., "What is Social Informatics and Why Does it Matter?," *D-Lib Magazine*, 5(1), 1999.
13. Doctors Without Borders/Médecins Sans Frontières (MSF); <http://www.doctorswithoutborders.org/>.
14. Engineers Without Borders – International; <http://www.ewb-international.org/>.
15. Gurstein, M., "E-commerce and community economic development: Enemy or ally?." SD Dimensions. FAO, Rome, 2000; <http://www.fao.org/sd/CDdirect/CDre0055i.htm>.
16. Gurstein, M., "Community informatics: Current status and future prospects." *Community Technology Review*, Winter-Spring, 2002.; <http://www.comtechreview.org>.

17. McIver, Jr., W. J., Elmagarmid, A. K. (eds) *Advances in Digital Government: Technology, Human Factors, and Policy*, Boston: Kluwer, May 2002.
18. International Telecommunication Union (ITU), *WORLD TELECOMMUNICATION/ICT INDICATORS*, ITU, 2006 May; http://www.itu.int/ITU-D/ict/statistics/at_glance/basic05.pdf .
19. Nielsen/NetRatings. "ONLINE SEARCHES GROW 55 PERCENT YEAR-OVER-YEAR TO NEARLY 5.1 BILLION SEARCHES IN DECEMBER 2005, ACCORDING TO NIELSEN/NETRATINGS." *NetRatings, Inc. Press Release*, 2006 Feb 9.; <http://www.nielsen-netratings.com> .
20. Organisation for Economic Co-operation and Development (OECD), "Data -- ICT patents as a percentage of total national patents filed at the EPO, for priority years 1990, 1998," OECD, 2006 Oct 23; <http://www.oecd.org/dataoecd/45/37/2766453.xls> .
21. Organisation for Economic Co-operation and Development (OECD), "11a. ICT patents as a percentage of national total (EPO) in selected countries," OECD Key ICT Indicators, Indicator 11a., 2006.; Organisation for Economic Co-operation and Development (OECD), "11a. ICT patents as a percentage of national total (EPO) in selected countries," OECD Key ICT Indicators, Indicator 11a., 2006.; Organisation for Economic Co-operation and Development (OECD), "11a. ICT patents as a percentage of national total (EPO) in selected countries," OECD Key ICT Indicators, Indicator 11a., 2006.; .
22. Organisation for Economic Co-operation and Development (OECD), "15. Contributions of ICT investment to GDP growth, 1990-95 and 1995-2003 (1), in percentage points," OECD Key ICT Indicators, Indicator 15., October 26, 2005.; .
23. San Diego Supercomputer Center (SDSC), "Grand Challenge Equations", University of California, San Diego, 1999; <http://www.sdsc.edu/GCequations> .
24. Association for Computing Machinery (ACM), *Computing Curricula 2005*, 30 September 2005; <http://www.acm.org/education/curricula.html> .
25. Green WiFi; <http://www.green-wifi.org/> .
26. Ó Siochrú, S., Girard, B., *Community-based Networks and Innovative Technologies: New models to serve and empower the poor*, A report for the United Nations Development Programme, 2005; http://propoor-ict.net/content/pdfs/Community_Nets.pdf .
27. Hammond, A., Paul, J., *A New Model for Rural Connectivity*, World Resources Institute, May 2006.
28. McDonough, W., Braungart, M., *Cradle to Cradle: Remaking the Way We Make Things*, North Point Press, 2002.
29. Broersma, M., "Spanish region goes entirely open source," *TechWorld*, 01 August 2006; <http://www.techworld.com/applications/news/index.cfm?newsID=6558> .
30. Global Text Project; <http://globaltext.org/> .
31. Rodney, W., *How Europe Underdeveloped Africa, (revised edition)*. Howard Univ Press, Washington, DC, 1981.
32. Cabral, A., "National Liberation and Culture," *1970 Eduardo Mondlane Memorial Lecture*, Syracuse University, 1970. (Translated from the French by Maureen Webster.)
33. Powell, M., "Knowledge, culture and the internet in Africa: a challenge for political economists." *Review of African Political Economy*, No. 88, 2001, pp. 241–266.

Appendix. An Overview of Human Development Indicators

The United Nations (UN) classifies "extreme poverty" the condition of living on less than \$1 (US) per day. The United Nations Development Programme (UNDP) reported

that the global proportion of people living in extreme poverty in 2001 as 20.7%. The proportion in Sub-Saharan Africa in 2001 was 46.4% [1].

The UNDP estimates that over 100 million children of primary school age are not enrolled in school, over 42 million of these are in South Asia [1].

The UNDP reported that from 1980 to the present, the child mortality in Sub-Saharan Africa increased from 13 to 29 times the rate of developed countries [1].

The UNDP estimates that over 1 billion people lacked access to clean water, 55 million of these are in Latin America and the Caribbean [1]. The World Water Council reports that in many regions around the globe increases in population, contamination, agricultural practices, and political conflicts continue in combination to create or heighten water scarcity [2]. The World Health Organization (WHO) reported in 2004 that 1.8 million people die annually because of lack of access to clean water [3].

The United Nations Programme on HIV/AIDS (UNAIDS) estimated that 38.6 million people were living with HIV in 2005, up from 36.2 million in 2003 [4]. An estimated 4.1 million adults and children were newly infected by HIV in 2005, up from 3.9 million in 2003. The global prevalence rate of people living with HIV has leveled off at 1% between 2003 and 2005, but the epidemic continues to expand in Southern Africa [4].

HIV/AIDS must be viewed from the perspective of its current and future impacts, however. So significant are they that UNAIDS has predicted that societal changes brought about by HIV/AIDS and the inability to respond adequately will impede attainment of the MDGs. This prediction is based on expected reductions in population overall; deaths of professionals, such as doctors and teachers in particular; and socio-economic shifts that have resulted, such as reductions in productivity as able-bodied people are forced to care for family members who are sick. Population reductions of over 100 million by 2015 – the year set by the United Nations for its Millennium Development Goal targets -- are predicted for the top 60 countries impacted the most by the epidemic [4]. The UNAIDS “2006 Report on the global AIDS epidemic” shows that Sub-Saharan Africa and Asia are currently experiencing the most significant impacts. In Sub-Saharan Africa, deaths from AIDS are already significantly impacting access to education, medical care, and other facets of a welfare state assumed by most in the developed world, as teachers, doctors and nurses, and professionals are themselves infected.^{4, 5}

The UNDP estimates that over 50 million girls of primary school age are not enrolled in school, over 5 million in Arab countries. Women and girls are at greater risk than men for living in extreme poverty and women are at greater risk for becoming infected with HIV.

Armed inter-state and intra-state conflicts have occurred on grounds of ethnicity, religion, and competition over territory and resources. They have been the cause of sustained damage to social and political infrastructure, which amplify social crises such as those surveyed above. Genocide has been formally recognized in several of these wars (see [6] and [7]).

⁴ HIV InSite (<http://hivinsite.ucsf.edu/>) provides a comprehensive set of statistics sources dealing with HIV/AIDS.

⁵ Former United Nations Ambassador and Special UN Envoy on HIV/AIDS, Stephen Lewis paints a vivid picture of these impacts [5].

Communications Breakdown: Revisiting the Question of Information and Its Significance for Community Informatics Projects

William Tibben

University of Wollongong, Wollongong NSW 2522, Australia
wjt@uow.edu.au
<http://www.dsl.uow.edu.au/~wjt>

Abstract. The gap between those who understand the complexities of community requirements and the information technologists who can build the technologies represents a central focus of concern with Community Informatics (CI) research. This paper explores how different assumptions about the utility of information during this innovation process leads to poor communication between researchers and practitioners. Braman's four-part hierarchy is a useful vehicle to investigate this as she seeks to include a range of actors such as policy makers, technologists and community members. A number of case study examples are explored to illustrate the value of Braman's work for CI.

1 Introduction

Developing an effective relationship between technologists and those who have intimate knowledge of community requirements represents an important focus of research within the field of Community Informatics (CI). Dependent on this nexus is the planning and deployment of effective Information and Communication Technology (ICT) based interventions to promote equitable outcomes for individuals and communities that are marginalised by economic, cultural or other factors. The task of bringing together disparate groups in order to develop an effective response is fraught with many challenges that are in need of clarification.

In responding to this problem the paper engages a research agenda that has been effective in the analysis of information systems and their use by individuals and groups [1]. By exploring assumptions that are entailed within different definitions of the term "information" a number of areas leading to poor communication between planners and practitioners emerge. The paper develops this understanding by using Braman's [2] four-part *definitional hierarchy* of information to analyse a number of case studies within the Community Informatics domain. The motivation for developing this line of enquiry lies in recognition of the socioeconomic and political factors that influence information creation,

processing, flows and uses. From this analysis a number of implications are drawn for CI researchers and practitioners to consider.

The paper is organised in the following fashion. The paper begins by establishing a broader theoretical context to understand research and practice within the domain of CI. The analysis goes on to propose that one significant gap in communication when planning for CI projects is factored on differing assumptions about the definition and utility of information. This contention is tested using Braman's [2] four-part definition of information-as-resource; information-as-commodity; information-as-pattern; and information-as-constitutive force. Case studies from CI are considered as each of Braman's definitions is explained. The paper finishes with a discussion about the significance of this analysis for CI research and practice.

2 Telecommunications Research and Community Informatics

The communications landscape has changed dramatically over the past two decades. Increasing penetration and connectivity of telecommunications on a world scale has opened up possibilities for economic and social development. The increasing miniaturisation and capabilities of electronic components together with the development of non-proprietary Internet protocols has brought with it a trend in which increasing network intelligence is being devolved to consumers. The increasing participation of communities in standards settings [3] as well as growing popularity of online discussion forums on the topic of the digital divide reflects a greater degree of participation enabled by ICTs. While ICTs have always been considered a critical aspect to development a qualitative shift is occurring. As Shearman [4] describes, people have become information *makers* rather than just mere *chasers* of information.

The International Telecommunications Union (ITU) hosted World Summit of the Information Society (WSIS) reflect a common agreement among nations that ICTs represent an important vehicle for the delivery of better development outcomes for those marginalised by economic or social factors. Retrospective analysis of the ITU's first concerted effort to link telecommunications with development, the Maitland Commission, reveals a disappointing history with some notable highpoints. Milward-Oliver's [5] twenty year review of research since the Maitland Commission places central the thorny task of bringing together the technological with the social. Community Informatics, as a more recent arrival in this research space, gives this linkage greater clarity with its emphasis on the end-user in its *effective-use* mandate. The effective use of ICTs according to Gurstein [6] aims to

'support local economic development, social justice and political empowerment; ensuring access to education and health services; enabling local control of information production and distribution; and ensuring the survival and continuing vitality of indigenous cultures'.

One aspect that is given special emphasis is the link between technologists and those in community that will make use of ICTs. Accordingly, Gurstein [6] suggests that a dialogue should be established between system planners and end-users. This two-way relationship recognises the fact that both planners and end-users are working within the limitations of their own knowledge. On the one hand, researchers do not fully comprehend the local circumstances in which end-users reside. On the other hand, end-users do not necessarily have the knowledge and experience to guide the deployment of ICTs. To this extent there needs to be a judicious mix of *technology push* and *technology pull* where the process is situated firmly within the cultural context in which the new ICTs will be deployed.

The terms *technology push* and *technology pull* represent two counter points in the study of technology transfer where the former is considered a top-down approach while the later is bottom-up.¹ As James [8] explains, the tension between these two extremes is factored on minimising the costs of innovation. Technology transfer was considered a preferable development strategy for the reason that the copying of technologies was theoretically cheaper for a country than independent development. As pointed out by economic historian Rosenberg [9] three decades ago, this idea is a simplistic one as technology transfer is rendered less effective by the absence of complementary technologies as well as complementary sources of knowledge. Such is the complexity of the process Hill [10] states that technologies are culturally bound and dependent suggesting that questions of technology transfer are ultimately about cultural change. In short, the costs of integrating technology into unfamiliar environments can be high leading to the conclusion that localised bottom up development is a necessary component of technology transfer.

In a CI setting it is not difficult to identify specific examples that support these observations of technology transfer. Absent complimentary technologies includes power and buildings while absent complimentary knowledge sources include IT specialists, multi-media experts and so on. The issue of the cultural divide between developers of technology and recipients represent a fundamental theme of the CI research agenda.

Further investigation of innovation research reveals that there are those who look to the study of information as a useful perspective on innovation and technology transfer. Macdonald [11] argues that the innovation process can be distilled into a common substrate of information-related concepts. Indeed, he details ways in which the innovation process is constrained and facilitated by the behaviour of people as they deal with complexities of information. Drawing on the work of Arrow [12] Macdonald maintains that the process of effectively responding to perceived needs, the development of plans, the putting together of

¹ In a theoretical sense, the *technology push* thesis holds that invention is the powerhouse of innovation as argued by Schumpeter while *technology pull* claims that the market leads innovation by demand for certain products through investment, as argued by Schmookler [7].

artefacts and the coordination of each can be understood in terms of the unusual economic characteristics of information.²

In contrast, information often appears as a relatively benign actor in the process of innovation especially if one considers the vast amounts of information that are available through the Internet and other sources. Machlup and al. [13] warn that the apparent homonymy of the word information may in itself present a barrier to deeper understanding. They state:

‘[i]nformation is not just one thing. It means different things to those who expound its characteristics, properties elements, techniques, functions, dimensions, and connections. Evidently there should be something that all things called information have in common, but it surely is not easy to find out whether it is much more than the name (p. 4)’.

Machlup and al. go on to describe the rarity of individuals who are able to span just some of the boundaries that separate the ‘30 or 40 cultures’ that represent information-related disciplines.

So it is with this apparent contradiction between the simplicity of the term “information” and its multi-disciplinary underpinnings that planners and practitioners in CI venture into partnerships to develop projects that have information as the fundamental resource being produced, stored, transferred and used. The potential for confusion appears obvious. The next section details one method by which the multifarious nature of information can be addressed.

3 Information: A Definitional Hierarchy

Braman’s [2] concern for the effective incorporation of information into public policy led her to develop four definitions for information. Braman’s four-part definitional hierarchy is as follows: information-as-resource; information-as-commodity; information-as-pattern and information-as-constitutive force. As an analytical tool each definition has varying degrees of utility because each entails a number of assumptions. Accordingly, ignorance of these same assumptions can potentially complicate the delivery of CI projects.

3.1 Information as Resource

The resource definition for information is one that Braman observes has widespread acceptance (pp. 235-236). As people are continually faced with the task of making decisions there is an intrinsic recognition that life can be improved by achieving greater access to information. Braman states that information in this context is judged to be akin to a physical economic resource. This leads to the adoption of economic measures that privilege the tangible over the intangible. An unfortunate side effect is that ICTs can then become a proxy for information

² The fundamental economic characteristics of information defined by Arrow [12] are its high fixed cost (is expensive to produce) and its low marginal cost (is cheap to copy).

leading to the view that the mere presence or absence of ICTs becomes the primary yard stick by which outcomes are measured. As a consequence it is difficult to engage effectively with the question of providing *useful* information because the project objectives are couched in the excessively vague notion of information access.

3.2 Information as Commodity

The second definition of information that Braman defines is the one of commodity (pp. 236-238). This definition of information recognises that information can be bought and sold. It also has the advantage of distinguishing information flows and the way these flows create economic value. It can be seen that the thesis supporting this concept is more refined than the previous definition in that it provides a clearer rationale for system developers to follow. Internet commerce is dependant upon information-as-commodity and as a consequence this aspect of information use has been an important focus of activity in the e-commerce domain. However, this definition of information has its limitations because it excludes information that may not be amenable to commoditisation.

3.3 Information as Pattern

The third definition of Braman's, information-as-pattern, introduces researchers and practitioners to the close relationship that exists between useful information and communities (pp.238-239). In short, information-as-pattern seeks to address the importance of context when managing information. This definition asserts that complementarities must exist between new information and existing information in order for the former to become productive. Existing information can take the form of individuals with requisite capabilities, cultural traditions as well as ICTs that give access to relevant sources of information and tools to configure information. The disadvantage of the information-as-pattern approach is its relativistic nature where the value of information to an individual or a group may be difficult to specify in advance.

3.4 Information as Constitutive Force

The final category, information-as-constitutive force, represents the pre-eminent definition from Braman's perspective (pp. 239-241). This view not only recognises that information is context dependent but accords information with agency to bring about change in communities. This definition alludes to the productivity of information where the provision of useful information leads to significant social benefit because information reproduction is very cheap. Alternatively, some information may undermine established norms and create discord within communities. Braman alerts readers to the issue of power as an influential factor that shapes agreements about "useful" information and the nature of change that such information will engender.

4 Analysis of Case Studies Using Braman's Definitional Hierarchy

Having described Braman's hierarchy of definitions for information it is now possible to use this framework to analyse selected case study examples. The following descriptions and analysis is designed to establish a deeper understanding of information-related factors that can complicate the delivery of CI projects.

4.1 Information as Resource

As previously stated the resource definition for information enjoys widespread acceptance. Information in this context is judged to be akin to a physical resource. This in turn leads to the adoption of economic measures that privilege the tangible over the intangible. Braman claims that ICTs can then become a proxy for information leading to the view that the mere presence or absence of ICTs becomes the primary yard stick by which outcomes are measured.

This simplistic rationale is sometimes used to justify the provision of ICT-based initiatives. One example of this can be seen in the Australian experience with the establishment of government funded Community Technology Centres (CTCs). This project was funded under a limited term funding arrangement that ended in July 2005 called "Networking the Nation" [14]. A focus on infrastructure can be seen in the way a user-pays regime was instituted to mediate access to equipment such as computers, printers and video conferencing facilities. The time frames in which CTCs were to gain financial independence has proved to be too short and this has left many CTCs struggling to keep their doors open. Insufficient definition of information needs special to local communities has led to a poor appreciation of the role these CTCs play which in turn has precluded serious consideration of further public subsidies [15] [16]. One significant ramification of reliance on this definition of information is that a prediction of Gurstein's [6] is likely to be realised where a second-class citizenry is institutionalised because the infrastructure effectively locks communities into linkages with ineffectual sources of information.

4.2 Information as Commodity

The information-as-commodity definition recognises that information can be bought and sold. This definition distinguishes flows within information value chains. Internet commerce is best explained using this definition so it comes as no surprise that some CI initiatives focus on e-commerce. For example, the People First Network in the Solomon Island represent a successful e-commerce initiative to overcome the distance that island communities must deal with in this country [17]. An innovative HF-based email system enables villagers to coordinate and place orders with distance suppliers. As ICTs can be clearly seen to improve the efficiency of commerce in a developing country setting information-as-commodity represents a plausible choice for system developers.

One significant limitation of this definition is that it excludes information that may not be amenable to commoditisation. As Lambertson [18] states ‘information can be a commodity but only to a limited extent (p. 25)’. In a similar vein, Gurstein [6] warns that CI projects may be sold short if understood uniformly as e-commerce ventures. One outcome is the promotion of consumerism where the end-user is perceived merely as a passive receiver of information and goods. In this context, the primary beneficiaries are the manufacturers who are keen to sell their digital wares into new markets sometimes facilitated by subsidies from public and private organisations.

However, there are examples where information-as-commodity can be used to strategically support equity goals. An example of this can be seen in the way a community newspaper has become an important source of revenue for one CTC. The CTC, located in Sussex Inlet on the south coast of New South Wales, Australia, has used its community connections and its printing equipment to produce a newspaper that is supported by advertising from local businesses. This revenue is used to subsidise other aspects of the CTC’s operation. This indicates an arrangement that can flexibly deliver equity outcomes while drawing on the commodity aspect of information.

4.3 Information as Pattern

Information-as-pattern establishes the importance of context thereby acknowledging the close relationship that exists between useful information and communities. This definition recognises that complementarities must exist between new information and existing information in order for the former to become productive. Such complementarities come to the fore in the following examples.

The UNESCO-funded programme that uses ICTs to reduce poverty among minority groups in the Indian sub-continent provides an example where close consideration is given to existing culture mores, the uniqueness of local contexts and associated literacy skills. The use of video and audio production is found to be more effective with people whose dominant mode of cultural reproduction is by oral accounts and visual means. Slater and Tacchi [19] observe that radio and video production is more trusted and familiar in contrast to the Internet and computers. They claim that the most promising avenue for using computers and the Internet is incorporating these into multimedia mixes using radio and video.

This project raises questions about India’s success in exporting IT services. An information-as-pattern analysis is able to probe the apparent dichotomy between the \$20.3billion contribution [20] this makes to the economy and the poor payback to middle and lower income Indians in terms of IT education. While one influential strand of theory in development economics called “trickle-down economics” claims that the benefits of India’s export from the IT services industry should logically flow “down” to other members of society Srinivasan and al. [21] observe that this has not been translated into providing accessible IT education. Apparently, services supplied by Indian IT companies to their global markets are not easily translated into appropriate educational methods for

middle and lower income Indians. An information-as-pattern understanding allows one to see the highly contextualised nature of these information needs and the unique skills of individuals who can bridge India's existing IT capability with the information-related needs at the community level.

When viewed from a pattern perspective the relativistic nature of information makes it difficult to state unambiguously what specific information is required in a given situation. In response, researchers have sought to establish the legitimacy of concepts such as "social capital" to link the social good aspects of information provision presumably with the economic concept of capital [22]. Social capital seeks to establish the value of linkages and trust between people as an indication that worthwhile information is being generated and exchanged. In order to make this concept more concrete parallels are drawn with schools and libraries thereby establishing a rationale for Government support of CI projects [16].

Another consequence of the relativistic nature of valuable information is the proliferation of disciplines that make claim to expertise within the CI domain. Where before telecommunications research was dominated by engineers and economists, the study of CI incorporates a broad church ranging in areas from education, telemedicine, media studies, social work, information systems and so on. An added complication for CI practitioners emerges where communication across disciplinary divides can also constrain productive outcomes as well as the more fundamental challenge in establishing productive communication between system planners and end-users.

4.4 Information as Constitutive Force

The last of Braman's definitions, information-as-constitutive force, represents the most potent definition in terms of socioeconomic issues. Not only is information context dependent but such information can influence the direction and nature of change in communities. Within a CI context, information-as-constitutive force can be fundamentally understood as providing information that enables people to make decisions for the benefit of themselves and the community. One example of this can be seen in the provision of health information that brings about beneficial changes in the behaviour of people. Another example is the broadcast of a local weather report that helps a farmer make decisions about his priorities for the day.

On the other hand, there are circumstances in which CI projects seek to alter established power relationships within a community. The previously cited UNESCO-sponsored programme in the India subcontinent [19] has as one of its objectives 'ICT initiatives that change social norms that disadvantage marginalised people (p. 2)'. Despite the benefits of this programme, the report also acknowledges that such changes can have a down side in that 'power contests [can] threaten participants and initiatives (p. 90)'. The potential for disagreement and conflict is real.

The idea of externalities becomes relevant to this definition where the prospect of unintended consequences, positive or negative, extends beyond the immediate confines of a CI initiative. A more visible example of this can historically be seen

in public broadcasting in the Pacific region. As one of its ideals the promotion of free-speech through broadcasting is seen as important for a host of reasons that relate to political and economic development. This in turn has facilitated foreign aid spending for broadcasting. However, broadcasting has not engendered an appreciation of the need for a free press within the political leadership of many Pacific Island countries [23]. This brings into stark focus an ideological clash between project sponsors and recipients because of different worldviews. It is perhaps for this reason that one observer suggests that small-scale media technology centres are a safer choice where the likelihood of political interference is less [24].

The issue of power looms large when considering the influence information has over the direction and nature of change. It is possible to understand the development of exploitative relationships in a situation of unequal financial resources. A similar circumstance may exist in relation to imbalances in knowledge. Perhaps the word 'imbalance' is ill advised in that the relative weighting accorded to individuals' knowledge may be based on biased value judgements. For example, should knowledge about the function of a computer operating system be given more credence than the traditional knowledge of a community leader? Ideally the answer is 'no' but achieving the right environment in which an optimum mix of knowledge from stakeholders leads to interventions that strengthen target communities is complex and difficult.

5 Discussion and Conclusion

From the perspective of marginalised communities wishing to fully profit from new ICTs, decisions about technology ultimately appear to be about information and what information works best in each local context. Braman's hierarchy of definitions has the potential to orientate the thinking of those planning CI projects in four ways.

Firstly, it is unhelpful to maintain vague notions of information as a resource because it tends to promote a dichotomy based on the presence or absence of ICTs. Gurstein [6] warns that the provision of ICTs with insufficient understanding of local context tends to promote a second class of citizenry.

Secondly, the idea that information may be commoditised to create commercial opportunities is welcome but clearly must be understood within the broader information melee. A danger exists where information that has high social benefit but insufficient commercial potential is not communicated. A corollary to this situation is the established media of broadcasting and print who for many decades have been using commoditised information to sell advertising which, in turn, subsidises the provision of other information such as news and current affairs.

Thirdly, a complicated mix of contextual factors that includes individual capabilities, social structure and cultural practices drives the productivity of information in any given environment. As a consequence, there is a need to understand local epistemologies and ontologies and the gaps within as a means to

developing effective ICT related responses. This has engendered innovative research agendas in a range of disciplines that look to the promise of multi-media technologies and the Internet for each of their area's challenges in overcoming marginalisation. High on the list of essential attributes for individuals is a multi-lingual capability that enables boundary spanning between disciplines as well as cultures. According to Matchlup and el. [13] such individuals are rare.

Fourthly, and building on the previous point, is the potentially transformative effect ICTs may have in a given situation or conversely, the disruptive effect as local customs and patterns of communication are altered. As Slater and Tacchi [14] advise, the highly uncertain nature of this activity suggests the need for iterative and evolutionary development and deployment of CI projects that are guided by contributions from a broad range of actors over a long period of time.

In summary, it appears that the task of system builders in the Community Informatics domain can be related in many ways to the vagaries of information - the manner in which it is defined and the factors that govern its flow and uses within communities. In describing the task of researchers in information Simon [cited in 13] likens the project to anthropology.

'We go into areas whose inhabitants speak foreign languages (with many words sounding like words in our language but have very different meanings); we try to find some guides to help us learn the meanings of strange sounds; and we try to make sense of what we see and hear, yet we probably misunderstand much and are bewildered by even much more (p. 5)'.

Such a description appears fitting also for those working at the coal-face in Community Informatics.

References

1. Lamberton, D. M. (ed.): *The Economics of Communication and Information*. Edward Elgar Publishing Company, Cheltenham, UK (1996)
2. Braman, S.: *Defining information*. *Telecommunications Policy*, Vol. 13(3), (1989) 233-242
3. Given, J. and Goggin, G. (eds.): *Choosing Telecommunications? Consumers in a Liberalised Privatised Telecommunications Sector*. Media International Australia, Vol 96 (2000) 1-134
4. Shearman, C.: *Strategies for reconnecting communities: creative use of ICTs for social and economic transformation*. In: Marshall, S., Wallace, T. Yu, X. (eds.): *Closing the digital divide: transforming regional economies and communities with information technology*. Praeger, Westport (2003) 13-26
5. Milward-Oliver, G.: *Maitland+20: Fixing the Missing Link*. The Anima Centre, Bradford-on-Avon (2005)
6. Gurstein, M.: *Effective use: A community informatics strategy beyond the digital divide*. *First Monday*, Vol. 8(12), (2003) Available on <http://www.firstmonday.dk/issues/issue8.12/gurstein>
7. Bijker, W.: *Sociohistorical technology studies*. In Jasonoff, S. Markle, G. Peterson J. and Pinch, T. (eds.): *Handbook of Science and Technology Studies*. SAGE Publications, Thousand Oaks, California (1995) 229-256

8. James, J.: *Information Technology and Development: a New Paradigm for Delivering the Internet to Rural Areas in Developing Countries*. Routledge, London (2004)
9. Rosenberg, N.: Economic development and technology transfer. *Technology and Culture*, Vol. 11(4) (1970) 550-575
10. Hill, S.: *The Tragedy of Technology: Human Liberation Versus Domination in the Late Twentieth Century*. Pluto Press, London (1988)
11. Macdonald, S.: Technology beyond machines. In Macdonald, S. and Lamberton, D. and Mandeville, T.: *The trouble with technology*. Francis Pinter, London (1983) 26-36
12. Arrow, K.: Economic welfare and the allocation of resources for invention. In Nelson, E.: *The Rate and Direction of Inventive Activity: Economic and Social Factors*. The Universities-National Bureau Committee for Economic Research, Arno Press, New York (1962) 609-626
13. Machlup, F. and Mansfield, U. (eds): *The Study of information: interdisciplinary messages*. Wiley, New York (1983)
14. NSW Department of Commerce: *Joint Commonwealth and New South Wales Community Technology Program: Final Project Report*. NSW Department of Commerce, Sydney. (2004)
15. Geiselhart, K.: *The electronic canary: sustainability solutions for Australian tele-service centres*. Community Teleservices Inc. A report commissioned by the Department of Communications, Information Technology and the Arts, Australian Government, Wangaratta (2004)
16. Simpson, L. Daws, L. and Pini, B.: Public internet access revisited. *Telecommunications Policy*, Vol.28 (2004) 323-337
17. Chand, A. Leeming, D. Stork, E. Agasi, A, and Biliki, R. : *The impact of ICT on rural development in the Solomon Islands: the PFNet case*. University of the South Pacific: Suva, Fiji (2005)
18. Lamberton, D.M.: *The Information Economy revisited*. In Babe R. (ed.): *Information and Communication in Economics*, Kluwer Academic Publishers, Dordrecht (1994) 1-33
19. Slater, D. and Tacchi, J.: *Research: ICT Innovations for Poverty Reduction*. UNESCO (Asia Pacific Regional Bureau for Communication and Information), Delhi (2004)
20. Mattoo, A. Mishra, D. and Shingal, A.: *Sustaining India's services revolution: access to foreign markets, domestic reform and International negotiations*. The World Bank, New Delhi (2004)
21. Srinivasan, A. and Awasthi, P.: *Building a knowledge society: the role of the informal sector*. In ICEG 2003 - International Conference on E-Governance. India Institute of Technology, New Delhi (2003)
22. Simpson, L.: *Community Informatics and sustainability: why social capital matters*. *The Journal of Community Informatics*, Vol. 1(2) (2005) 79-96.
23. Seward, R.: *Radio Happy Isles: Media and Politics at Play in the Pacific*. University of Hawai'i Press, Honolulu (1999)
24. Molnar, H.: *Communication technology: in whose interests?* In Marjoram, A. (ed.): *Island Technology: Technology for Development in the South Pacific*, Intermediate Technology Publications Ltd, London (1994) 104-117

More Than Wires, Pipes and Ducts: Some Lessons from Grassroots Networked Communities and Master-Planned Neighbourhoods

Mark Gaved¹ and Marcus Foth²

¹ Knowledge Media Institute, The Open University
Walton Hall, Milton Keynes, MK7 6AA, United Kingdom
m.b.gaved@open.ac.uk

² Institute for Creative Industries and Innovation, Queensland University of Technology
Creative Industries Precinct, Musk Ave, Brisbane QLD 4059, Australia
m.foth@qut.edu.au

Abstract. Community informatics research has found that the provision of technical connectivity in local neighbourhoods alone does not ensure community interaction. Externally initiated projects applied to communities by government or commercial bodies have encountered difficulties where the project's goals do not correspond to the host community's. Differing expectations can lead to disillusionment or rejection. Self-organised initiatives developed from within communities appear to be more aligned with residents' goals and purposes and may not face these issues. However, such initiatives have also encountered difficulties in maintaining volunteer input and achieving technological sustainability. Valuable insights can be drawn from both cases. In this paper we review examples of each type of initiative and consider lessons that can be taken forward into new networked neighbourhood initiatives currently being developed. We consider one specific example, an inner-city master-planned residential development in Australia seeking to establish a community association to support socio-economic sustainability and governance of the local ICT infrastructure. We offer recommendations drawn from existing projects that may be applied to this site and to a wider context, and consider some implications for the future selection, deployment and maintenance of community information systems.

Keywords: community informatics; community information systems; community networks; grassroots communities; networked communities; master-planned communities; urban neighbourhoods; ICT.

1 Introduction

Technological solutions to facilitate social communication and interaction between residents in urban populations are increasingly important for many countries. In this paper we consider Australia and the UK which are particularly prone to issues that stem from regional migration, socio-cultural diversity, urban renewal and increasing densification. Such trends are also evident in other urbanised areas around the world.

Both the private and public sectors are looking at information and communication technology (ICT) to take on a mediating role to remedy some of these issues.

Australia and the UK are two of the most urbanised countries in the world in terms of the high proportion of urban dwellers among its total population. The increased population density generates rising demand for ICT infrastructure and services to enable social communication and interaction between urban residents. Internet cafés are a familiar sight in urban centres, mobile phone reception and wireless Internet services are approaching near full coverage of inner-city areas. Surprisingly, despite the fact that urban centres are possibly better connected than they have ever been before, notions of urban alienation are still being discussed in urban planning and policy making. Randolph (2004, p. 483) argues that, *“the language of community has come back with vengeance in policy areas that ignored it for many years. Cities are becoming, perhaps more than ever before, collections of distinctive communities and neighbourhoods, all the more differentiated as the cities grow in size and complexity. As the city expands, people remain focused on their small part of it.”*

Randolph's more contemporary image of community is consistent with Wellman's (2001) notion of networked individualism which reflects the inherent duality of the collective group networked by means of ICT and the individual who wants to stay in control of their social circle – their ‘small part of the city’. Social isolation and ‘non-connectedness’ have high social and economic costs (DCITA, 2005). ICTs that enable the formation of community networks can help bring collocated people together socially and increase awareness of individual and community skills and assets. Community networks can also support access to local information and services as well as encourage public consultation and open debate

Scholars such as Gurstein (2003) argue that the provision of access itself is necessary but not sufficient to ensure meaningful usage. Furthermore, providing connectivity and electronic access by setting up online portals and discussion boards does not automatically ensure user participation and the development of a sense of community. However, many externally driven ICT initiatives (e.g. funded by government or commercial organisations) have built tools and infrastructures with the assumption that ‘if you build it, they will come’. This approach assumes a local willingness to participate. Although these websites may provide useful community information and services, the ability of such systems to connect residents with each other can be limited. They may be seen as ‘yet another burden’ instead of a helpful communication tool which can be used to support the social networks which residents form with friends and peers.

An alternative approach has been the development of grassroots, self-organised initiatives driven by existing communities. Local activists have constructed their own network infrastructure with minimal external support to enhance local communication and a sense of community. These may be more successful in encouraging take up and usage, as increased local control and management fosters community ownership. Yet technical expertise is harder to maintain within a small community, and thus the social sustainability of a network run on little or no funds and dependent on volunteers may prove to be problematic.

In this paper we review externally initiated (‘top-down’ or ‘master-planned’) and self-organised (‘bottom-up’ or ‘grassroots’) projects and gather some lessons to inform the development of new community networks. Findings are set against the

emerging requirements of an inner-city master-planned residential development in Australia seeking to establish a community association to support the medium and long-term economic and social sustainability and governance of the local ICT infrastructure.

2 Externally Initiated Community Networks

University researchers have been involved in developing and supporting community networks since the 1970s and the Berkeley Community Memory (Farrington & Pine, 1996). Governmental enthusiasm to connect society and industry to the new 'communication superhighways' (Hearn *et al.*, 1998) caused concerns to be raised about a possible divide in access to ICTs. This led to research in the form of widespread surveys and pilot projects to explore the effects of supporting connectivity. Key projects have included Netville in Canada (Hampton & Wellman, 2003), Blacksburg Electronic Village (Cohill & Kavanaugh, 2000) and Camfield Estates (Pinkett, 2003) in the USA, Ennis in Ireland (McQuillan, 2000), Williams Bay (Arnold *et al.*, 2003) and Atherton Gardens (Hopkins, 2005) in Australia and the Wired Up Communities in the UK (Devins *et al.*, 2003). These projects have operated across a broad range of cultures and with a wide variety of circumstances, but common issues have arisen.

Many externally initiated community networks have been run with a fixed timeline: actions are undertaken, data is collected and the project written up. In some cases the participants are aware of this limitation from the outset; for example by being offered free internet connectivity for a set period of time (Devins *et al.*, 2003). In other cases this came as a complete surprise – in Netville, residents assumed their internet access was permanent as part of their house purchase and were upset when the technology consortium announced it had gathered its data and would be closing down the service (Hampton, 2003). Other projects may struggle to survive after external funding has finished and are forced to change their priorities and targets as a result leading to a failure to support the original clients (Hopkins, 2005).

Such a 'project based' approach to community networks may make them less socially sustainable – users may be encouraged to participate but are then left unsupported and disenfranchised as a result. Day & Cupidi (2004) recommend that community technologies should be approached as open ended initiatives rather than closed term projects, as the latter is detrimental to social sustainability. If a community network is to be of long term benefit it must be seen as part of the long term infrastructure and strategies. Exit strategies must be formulated to ensure the network can continue after funding has finished; these should consider not only infrastructure funding but also community support, training, and staffing.

Externally initiated projects may suffer if they do not consider local social structures; while this is more significant in existing communities, new communities also rapidly develop social structures and these must be taken into account. In the Ennis 'Information Age Town' project (McQuillan, 2000) a wide range of technological applications were put into place alongside the ICT infrastructure. However in some cases these destabilised rather than developed social cohesion. For example, unemployed people were asked to sign up for their unemployment benefits

online rather than by visiting the town's unemployment centre. While this may have sped up a clerical process, it removed an important social ritual for already isolated individuals within the community (Warschauer, 2002). Similarly, low usage of a telecentre in the Wired Up Communities project (Devins *et al.*, 2003) was later found to be due to its physical location in a community venue that had been the central meeting point during a recent coal miners' strike. Such a politically charged venue would not be used by a significant number of the local residents for this historical reason. Both examples illustrate the need to consider the wider communicative ecology of the community and locale (Foth & Hearn, 2006).

Community networks are technologically complex; they offer services to a local neighbourhood that are comparable to a business IT department. In this aspect, externally initiated community networks often perform well. Budget is allocated for set up and support of the explicit aspects of the network – the devices and the ICT infrastructure itself. Association with university technology researchers can assure free or low cost technical support (e.g. Cohill & Kavanaugh, 2000). However such resources need to be available on a long-term open-ended basis, with strategies for continued support if the project has limited time period funding.

Projects may encounter 'social resistance' with not all members of a locality interested in signing up to join the community network (Selwyn, 2003). This will have an effect on any network that seeks to be inclusive and providing a medium for all residents of a locality. Externally initiated networks particularly have encountered difficulties of being perceived as being driven by external goals not relevant to the local community. This may lead to members of the locality not being connected, or disconnecting, thus reducing the social effectiveness of such a network.

3 Self-organised Community Networks

Since the earliest days of the internet, there have been grassroots driven local community based applications of network tools and services. Many of the early bulletin board systems and Free-nets were started by innovators within local communities to support neighbourhood activities (Schuler, 1996) and this spirit has continued until the present. Similar to other earlier technologies such as the radio and the telephone, innovators and early adopters within communities have appropriated new ICTs for their own uses, either as a response to lack of provision, adapting inflexible exogenous framings of technology, or innovating for their own purposes (Jankowski, 2006). These initiatives are often funded from within the community, are volunteer run and respond to the demands of the local population (Gaved & Anderson, 2006). While they may be set up and run for a variety of motivations, their initiators often claim that the self-organisation makes them more sustainable in the long term than externally initiated projects: *"the very fact that the project is not dependent on external money means that there is nothing to run out of"* (Davies, 2004, p. 46). In many cases the funding model is more complex, with partnerships and reciprocal relationships supporting local activity, however there is usually an emphasis on local control and management.

Self-organised network communities driven from within local neighbourhoods have been less well researched than externally initiated networks (Gaved &

Mulholland, 2005), but it is clear that they are not homogenous, and offer lessons that can be carried forward. Such initiatives are usually seen as being more in touch with local community aspirations and goals. As the organisers are from within the community, ownership is more likely to be seen as being held by and more accountable to the residents. Furthermore, ongoing local support and training are considered of great importance to sustain the network.

While such initiatives often seek complete coverage within the neighbourhood, there will be non-participants (Selwyn, 2003). In addition, multiple subcultures are found within self-organised network communities, and it is likely that different groupings exist within a neighbourhood. Thus, features and tools that require an ongoing critical mass of users may prove to be more difficult to sustain than systems that connect with existing social networks and engage residents in a less homogeneous and more diversified manner (Foth, 2006a, 2006b).

All tools and services are likely to be appropriated and may not be used as designed, and there may be varying rates of success. It is likely that some tools provided within the network may not be used at all and an iterative implementation process is required. The Redbricks community network in Manchester started with a large variety of tools including music and video sharing (Skyva, 2002) but have reduced these services to two email lists: 'Shout' for calls to the whole community and 'Act' for political issues. Davies (2004) suggests that the most effective tools within a community network are those that offer non-critical services that help to build social capital, such as baby sitting services. Basic recommendation tools such as local noticeboards may be more effective than more complex services.

Self-organising network communities appear to be more socially sustainable due to their locally initiated nature, however they may struggle with financial and technical sustainability. Just as it is important to emphasise the value of the community network as a shared resource so it is important to recruit volunteers with the necessary skills and expertise to support and develop the network. Small networks may struggle to maintain the level of expertise required and benefit from participation in social networks of similar groups, for example the Community Broadband Network (www.broadband-uk.coop) in the UK (mainly focused on rural network groups), or NYCWireless (www.nycwireless.net) in the USA (aimed at wireless network groups in New York City and the surrounding areas).

It is clear that self-organised community networks have both strengths and weaknesses, as have externally initiated networks, and we now turn to consider how these findings may inform a new partnership based community network that is being developed in Australia.

4 Master-Planned Communities

Many new urban developments are systematically planned and rapidly built and marketed, trying to create instant 'communities' in dense concentrations. Developers and governments around the world struggling to achieve socially sustainable neighbourhood communities in these urban contexts, are increasingly considering the role of ICT to help animate master-planned communities (Foth, 2006d).

Gleeson (2004) gives examples of the prevailing attitude of developers who confuse 'planning for community' with 'master-planning community' and the associated negative impact on community development efforts. "*Community development involves human horticulture, rather than social engineering*" (Gilchrist, 2000, p. 269). The Kelvin Grove Urban Village (www.kgurbanvillage.com.au) is a master-planned residential development in inner-city Brisbane that seeks to learn from these and other lessons.

Queensland University of Technology and the Queensland Government's Department of Housing have established a partnership to develop the Kelvin Grove Urban Village (KGUV), an integrated master-planned urban renewal project. They have identified the KGUV as a distinct planning and design case study that departs from homogeneous planning principles. It reflects a desire to achieve a higher level of integration of population diversity than has been aimed for in past urban renewal initiatives (Healy & Birrell, 2004). One of the innovative aspects of the KGUV is the provision of housing types for a range of income groups, such as 'mainstream' apartments, senior and student accommodation, and affordable housing. The collocation of a diverse socio-demographic population within a new inner-city area offering a mix of residential, commercial, educational, cultural and employment facilities and activities will inform the objectives of the ICT strategy to support social cohesion and may influence uptake and usage.

The objective to create a vibrant place of mixed uses and diverse population is reflected in the KGUV vision statement: "*A diverse city fringe community linking learning with enterprise, creativity with community and unique living solutions with public amenity.*" Realising this bold vision requires a theoretically and empirically grounded understanding of how urban neighbourhoods can be assisted to emerge and grow in healthy ways through community development activities and the role of ICT to assist in this effort.

Research is underway to examine issues of socio-cultural sustainability in the experience of residents settling into a new environment. The Department of Housing breaks down the concept of urban sustainability into the 'triple bottom line' (Gleeson *et al.*, 2004) of environmental, economic and social sustainability. This research project focuses on the social component (Buys *et al.*, 2005) by engaging a tripartite approach comprising community capacity building strategies (*people*), a theory of neighbourhood identity based on 'networked individualism' (Wellman, 2001) (*place*), and design of online community networks (*technology*). These three components are inter-related. The study thus employs an inclusive approach that seeks to overcome any tendencies to ignore key factors in the design and development of meaningful ICT applications for residential communities.

The provision and implementation of the ICT infrastructure at the KGUV seeks to prepare the site to play an important part of Queensland's emerging knowledge economy. The Queensland Government (2005) recognises the potential of ICT to enable people to work where they choose to live, connecting them with the world, and encouraging intellectual growth. It wants the local network to help create opportunities to integrate work and home life through high-speed, global communication systems for both businesses and residents. Common service ICT ducts have been installed beneath the footpaths in the KGUV, giving the potential to offer residents, home workers and business operators' broadband access to the

Internet, high-speed transmission between local stakeholders of the KGUV, and high quality telephone and audiovisual services. A commercial provider has been contracted to ensure the long-term continuity of technical development of the ICT infrastructure across both terrestrial and wireless networks.

However, the KGUV project team has started to translate the lessons learnt from the studies referred to above into action. They have realized that it requires more than the provision and installation of wires, pipes and ducts to achieve a socially sustainable urban village community. We briefly outline three key strategies which seek to distinguish this initiative from the pitfalls of previous projects.

First, the provision of ICT systems and related services is designed with an exit strategy in mind right from the start to ensure the main financial assistance from the primary stakeholders is made continuously redundant over time. The vision of the KGUV as a smart neighbourhood and inclusive community is driven by a range of community development activities, and the KGUV Community Association is one of the key initiatives. It will be established by the Department of Housing and Queensland University of Technology. KGUV residents represent the main group of prospective members of the Community Association. Whilst the mission and business plan of the Community Association is distinct from the KGUV Principal Body Corporate, both entities are established to ensure the medium and long-term economic and social sustainability and governance of the KGUV. The Association will be a commercial entity which develops, markets and sells creative industries' services. The main asset of the Association will be the KGUV Community Portal which is currently being developed by a commercial web development company and which will eventually be maintained and managed by the Association.

Secondly, the theoretical and methodological frameworks underpinning the project's research and development are based on principles of inclusiveness. In order to avoid considering a newly provided community network system in isolation, KGUV invokes the concept of 'communicative ecology' which we define as a milieu of agents who are connected in various ways by various media making. This notion integrates the three dimensions of 'online and offline', 'global and local' and 'collective and networked' (Foth & Hearn, 2006). This more holistic model helps us better appreciate the dynamic inter-relationships between different communication technologies and between different social dimensions found in the interactions of KGUV residents. It informs the creation of gateways and interfaces between existing social networks and communication systems on the one hand and the new KGUV Community Portal as a local communication hub on the other. Furthermore, network action research (Foth, 2006c) is used as a project methodology to reciprocally inform research and practice and to encourage community members to become reflective practitioners who take up community ownership of the initiative.

Thirdly, the project group has recognised the need to not only ensure network access but also effective use of the network by residents and other stakeholders. The portal aims at facilitating community uptake of ICT by hosting entertainment and information content that encourages exploration of the ICT infrastructure available at the KGUV. Furthermore, the portal offers an outlet for self-published local content which is intended to provide an online mechanism to link the people and businesses at the KGUV and beyond. It is supposed to encourage participation in the KGUV by being not only a key information resource for the diverse mix of activities, programs

and facilities available, but also a communication hub. The portal will focus less on collective communication features such as discussion boards and more on peer-to-peer modes of interaction to reduce reliance on maintaining a critical mass of users. Such features can act as a springboard to animate interaction which may be continued through external applications and devices like email, instant messaging software and mobile phones. This approach places less pressure on an online space to try drawing all residents together collectively and satisfying all their social needs and purposes, which in itself it may not be able to achieve (Foth, 2006a, 2006b).

The research, design and development of the ICT component and social sustainability aspects of the KGUV started in early 2006 (Foth, 2006d; Foth & Adkins, 2006). Evaluation strategies as part of the action research cycles will show whether the three broad strategies and principles discussed above make a significant difference in achieving a sustainable community network for KGUV residents.

5 Conclusions

In this paper we have described a variety of types of community network, and it can be seen that both externally funded and self-organised networks have shown both advantages and weaknesses. We have attempted to combine lessons learnt from both bottom-up and top-down approaches towards community networks, and introduced a new top-down / bottom-up hybrid initiative. The Kelvin Grove Urban Village seeks to build on the insights gathered from both 'community' and 'informatics' disciplines, that is, community development and information systems design. We identify the following three key recommendations for KGUV, and future community networks:

- **Cultivate a sense of ownership:** Community networks that are felt to be part of the community's own assets are those that are best supported and most socially sustainable in the long term. We recommend connecting internal and external interests and resources through a theoretical framework and methodological approach which combines research and practice, considers existing and emerging local social structures, and encourages community members to act as co-investigators.
- **Simple, open ended tools are the most successful:** Highly complex tools may be little used and too alien to be domesticated by the community. Simple tools that allow informal social dialogue have proved to be more successful. Additionally, it is not unreasonable not to try to connect everyone with everyone else. Peer-to-peer modes of communication are more conducive to supporting interaction in place-based social networks than collective, broadcast-style tools alone which require a constant critical mass of users to maintain momentum.
- **Develop externally initiated networks with an exit strategy in mind:** All users require technical support at some stage. Encouraging peripheral participation through buddying new users with expert users, providing online community help boards, informal and formal training will enable ongoing usage of the service and develop technical and managerial staff.

Clearly further research is required. Hence, data gathered from KGUV will be valuable and reported in future papers. It is highly likely that more partnerships of this

kind will be developed (for example, the Oakgrove Millenium Community of 1850 wired houses in the UK, to be occupied from early 2007) and the experiences of such new urban networked communities are likely to inform both digital divide policy and community informatics research in the future.

Acknowledgements

Mark Gaved's PhD research has been funded by the Engineering and Physical Sciences Research Council of the UK. Dr Marcus Foth is the recipient of an Australian Postdoctoral Fellowship supported under the Australian Research Council's Discovery funding scheme (DP0663854). Further support has been received from the Queensland Government's Department of Housing. The authors would like to thank Aldo de Moor, Paul Mulholland, Elija Cassidy, Michael Gurstein and the anonymous reviewers for valuable comments on earlier versions of this paper.

References

- Arnold, M., Gibbs, M. R., & Wright, P. (2003). *Intranets and Local Community: 'Yes, an intranet is all very well, but do we still get free beer and a barbeque?'* Paper presented at the 1st International Conference on Communities and Technologies (C&T), Amsterdam.
- Buys, L., Barnett, K., Miller, E., & Bailey, C. (2005). Smart Housing and Social Sustainability: Learning from the Residents of Queensland's Research House. *Australian Journal of Emerging Technologies and Society*, 3(1), 43-45.
- Cohill, A. M., & Kavanaugh, A. L. (Eds.). (2000). *Community Networks: Lessons from Blacksburg, Virginia* (2nd edition ed.). Norwood: Artech House.
- Davies, W. (2004). *Proxcommunication: ICT and the local public realm*. London, UK: The Work Foundation.
- Day, P., & Cupidi, R. (2004, Sep 29 – Oct 1). *Building and Sustaining Healthy Communities: The symbiosis between community technology and community research*. Paper presented at the Community Informatics Research Network (CIRN) Conference, Prato, Italy.
- DCITA. (2005). *The role of ICT in building communities and social capital: a discussion paper*. Canberra, ACT: Dep. of Communications, Information Technology and the Arts.
- Devins, D., Darlow, A., Petrie, A., & Burden, T. (2003). *Connecting communities to the Internet: evaluation of the Wired Up Communities programme* (DFES report No. RR389). Leeds: Policy Research Institute, Leeds Metropolitan University.
- Farrington, C., & Pine, E. (1996). Community memory: A case study in community communication. In P. Agre & D. Schuler (Eds.), *Reinventing technology, rediscovering community*: Albex.
- Foth, M. (2006a). Analyzing the Factors Influencing the Successful Design and Uptake of Interactive Systems to Support Social Networks in Urban Neighborhoods. *International Journal of Technology and Human Interaction*, 2(2), 65 - 79.
- Foth, M. (2006b). Facilitating Social Networking in Inner-City Neighborhoods. *IEEE Computer*, 39(9).
- Foth, M. (2006c). Network Action Research. *Action Research*, 4(2), 205-226.
- Foth, M. (2006d). *Research to Inform the Design of Social Technology for Master-Planned Communities*. Paper presented at the 14th European Conference on Information Systems (ECIS), June 12-14 2006, Göteborg, Sweden.

- Foth, M., & Adkins, B. (2006). A Research Design to Build Effective Partnerships between City Planners, Developers, Government and Urban Neighbourhood Communities. *Journal of Community Informatics*, 2(3).
- Foth, M., & Hearn, G. (2006). Networked Individualism of Urban Residents: Discovering the Communicative Ecology in Inner-City Apartment Complexes. *Information, Communication & Society*, (forthcoming).
- Gaved, M., & Anderson, B. (2006). *The impact of local ICT initiatives on social capital and quality of life* (No. 2006-6). Colchester: University of Essex.
- Gaved, M., & Mulholland, P. (2005). Ubiquity from the bottom up: grassroots initiated networked communities. In M. Consalvo & K. O'Riordan (Eds.), *AoIR Internet Research Annual* (Vol. 3). New York, NY: Peter Lang.
- Gilchrist, A. (2000). The well-connected community: networking to the edge of chaos. *Community Development Journal*, 35(3), 264-275.
- Gleeson, B. (2004). Deprogramming Planning: Collaboration and Inclusion in New Urban Development. *Urban Policy and Research*, 22(3), 315-322.
- Gleeson, B., Darbas, T., & Lawson, S. (2004). Governance, Sustainability and Recent Australian Metropolitan Strategies: A Socio-theoretic Analysis. *Urban Policy and Research*, 22(4), 345-366.
- Gurstein, M. (2003). Effective use: a community informatics strategy beyond the digital divide. *First Monday*, 8(12).
- Hampton, K. (2003). Grieving for a lost network: collective action in a wired suburb. *The Information Society*, 19(5), 417-428.
- Hampton, K., & Wellman, B. (2003). Neighboring in Netville: How the Internet Supports Community and Social Capital in a Wired Suburb. *City and Community*, 2(4), 277-311.
- Healy, E., & Birrell, B. (2004). *Housing and Community in the Compact City* (Positioning Paper). Melbourne, VIC: Australian Housing and Urban Research Institute.
- Hearn, G., Mandeville, T. D., & Anthony, D. (1998). *The communication superhighway: social and economic change in the digital age*. Sydney: Allen & Unwin.
- Hopkins, L. (2005). Making a Community Network Sustainable: The Future of the Wired High Rise. *The Information Society*, 25(5), 379-384.
- Jankowski, N. W. (2006). Creating Community with Media: History, Theories and Scientific Investigations. In L. A. Lievrouw & S. Livingstone (Eds.), *Handbook of New Media* (2nd ed.). London: Sage.
- McQuillan, H. (2000). *Ennis Information Age Town: A Connected Community*. Ennis: eircomm.
- Pinkett, R. D. (2003). Community technology and community building: early results from the Creating Community Connections project. *The Information Society*, 19(5), 365-379.
- Queensland Government. (2005). *Smart Queensland: Smart State Strategy 2005-2015*. Brisbane, QLD: Department of the Premier and Cabinet.
- Randolph, B. (2004). The Changing Australian City: New Patterns, New Policies and New Research Needs. *Urban Policy and Research*, 22(4), 481-493.
- Schuler, D. (1996). *New Community Networks: Wired for Change*. New York: ACM Press.
- Selwyn, N. (2003). Apart from technology: understanding people's non-use of information and communication technologies in everyday life. *Technology in Society*, 25(1), 99-116.
- Skyva, R. (2002). *The Hulme job: Redbricks Online Community Network*. Unpublished MSc thesis, Salford University, Salford.
- Warschauer, M. (2002). Reconceptualizing the digital divide. *First Monday*, 7(7).
- Wellman, B. (2001). Physical Place and Cyberplace: The Rise of Personalized Networking. *International Journal of Urban and Regional Research*, 25(2), 227-252.

Community-Driven Ontology Evolution Based on Folksonomies

Domenico Gendarmi and Filippo Lanubile

University of Bari, Dipartimento di Informatica, Via E. Orabona, 4 – 70125, Bari, Italy
{gendarmi, lanubile}@di.uniba.it

Abstract. The Semantic Web mission is to enable a better organization of the Web content to improve the searching, navigation and integration of the available information. Although the Semantic Web is intended for machines, the process of creating and maintaining it is a social one: only people, for example, have necessary skills to create and maintain ontologies. While most existing ontologies are designed by single individuals or small groups of experts, actual ontology users are not involved in the development process. Such an individual approach in creating ontologies, lead to a weak community grounding. On the other hand, Social Software is becoming increasingly popular among web users, giving opportunities to exploit the potential of collaboration within a community. Tools like wikis and folksonomies allow users to easily create new content and share contributions over a social network. Social Software tools can go beyond their current limits, by exploiting the power provided by semantic technologies. Conversely, Semantic Web tools can benefit from the ability of Social Software in fostering collaboration among users, by lowering entry barriers. In this paper we propose a new approach for ontology evolution, considering collaborative tagging systems as an opportunity to complement classic approaches used in maintaining ontologies.

Keywords: folksonomy, collaborative tagging, ontology evolution, social software, semantic web, semantic collaboration.

1 Introduction

The Semantic Web mission is to enable a better organization of the Web content to improve the searching, navigation and integration of the available information [2]. Current Semantic Web tools require a significant expertise level from their users, specifically in languages and techniques for knowledge representation. Other than being thought for computer science graduates, these tools do not provide any support for collaborative work.

Since Tim Berners-Lee's vision, online communities have taken an active role in the task of knowledge contribution on the Web. The recent phenomenon of Web 2.0 [13], also known as Social Software, has led to the growth of several tools which have succeeded in making this task more attractive to a broader audience. Social Software users are more than just passive information consumers but active participants, working in close collaboration to create new content and share it on the Web.

There are interesting research challenges at the intersection of Social Software and Semantic Web. Social Software can overcome its current limitations, by exploiting the power provided by semantic technologies in searching, navigation and integration of the information published on the Web. Semantic Web can benefit from the ability of Social Software in fostering collaboration among users, then lowering entry barriers to knowledge management.

In this paper we begin to investigate these opportunities, by presenting an approach which aims to combine the main benefits resulting from a specific type of social software, folksonomies, and well-defined formalisms applied in classical knowledge engineering techniques.

In the next section we discuss the drawbacks in knowledge creation and sharing. Section 3 introduces background information related to folksonomies, while Section 4 provides a short overview of known approaches to combine folksonomies and semantic techniques. Finally, Section 5 presents our approach and draws research directions.

2 Issues in Knowledge Sharing

Ontologies play a central role in the Semantic Web vision because they establish common vocabularies and semantic interpretations of terms accessible by machines. Sharing a common understanding of the structure of information among people and software agents is one of the main reasons for using ontologies [7].

Although the Semantic Web is intended for machines, the process of creating and maintaining knowledge is human-intensive. While most existing ontologies are designed by individuals or small groups of experts, ontology users are not involved in the development process. Such a restrictive approach in creating ontologies, leads to a weak community grounding.

The achievement of a widespread participation and shared consensus among ontology users is hampered by entry barriers, like the lack of easy-to-use and intuitive tools capable to include users in the ontology development process. The Ontology Web Language (OWL), for example, has the major drawback to be in several aspects non-intuitive for people who are not familiar with the Description Logics field.

Another relevant problem is the temporal extent of reliable knowledge which tends to be short. More information users learn, more the agreement and consensus among them evolve; thus new pieces of knowledge have to be committed and older pieces have to be constantly checked and validated. However, current ontologies require that all the changes have to be captured and introduced by the same knowledge engineers who created them. The delay in realizing and introducing changes can take weeks or even months, a period of time unacceptable in many dynamic domains, where knowledge regularly and rapidly changes. To be really effective, ontologies thus need to change as fast as the parts of the world they describe [8].

3 Collaborative Tagging

The term folksonomy, a combination of the words “folk” and “taxonomy”, is meant to indicate the act of collaboratively tag resources within Internet communities [16].

Personal user's metadata, commonly called tags, can be applied to any public resource and shared among all the community participants. A folksonomy, also known as collaborative tagging system, exploits the classification activity performed by each user, creating a network of users, resources and tags with a flat structure and no limits in evolution.

People have been starting to legally publish and share their content on the web, in the form of bookmarks, academic paper references, pictures, short audio and video files. Services like del.icio.us, CiteULike, Flickr, Last.fm and YouTube empower users to organize all these information resources through simple keywords.

Keyword-based systems for organizing digital content have already existed for many years. Digital libraries use metadata to discriminate between relevant and irrelevant information needs. While metadata have been long-used to improve information organization and discovery, they are usually created by knowledge engineers or by the original authors [10].

The key element underlying the proliferation of collaborative tagging system is the opportunity for users to collaborate in categorizing all the available resources with no forced restrictions on the allowed terms. In addition to individual benefits coming from the lack of vocabulary restrictions, the whole community can achieve significant advantages resulting from each participant contribution.

Furthermore, collaborative tagging systems create a strong sense of community among their users. Users are able to realize how others have categorized the same resource or how the same tag has been used to label different resources. This immediate feedback leads to an attractive form of asynchronous communication through metadata. Marginal opinions can coexist with popular ones without disrupting the implicit emerging consensus on the meaning of the terms rising up the folksonomy.

Opposed to professionally developed taxonomies also called controlled vocabularies, folksonomies show interesting potential to overcome the vocabulary problem [4]. One of the major obstacles hindering the widespread adoption of controlled vocabularies is the constant growth of available content which anticipates the ability of any single authority to create and index metadata [11]. While professionally created metadata are characterized by high quality, they do not scale up. On the contrary, collaborative tagging systems have a very short learning curve because there is not a predefined structure and syntax to learn.

The most relevant strength of a folksonomy is its ability in adhering to the personal way of thinking. There is no need for establishing a common agreement on the meaning of a tag because it gradually emerges with the use of the system. Folksonomies can then react quickly to changes and be responsive to new user needs.

4 Hybrid Folksonomy-Based Approaches

Folksonomies and ontologies have been often seen as competitors, the hype generated by collaborative tagging has inspired long debates over the web about which can be the best approach to categorization [12], [15]. Recently the debate has moved towards scholarly literature, analysing collaborative tagging phenomenon more rigorously.

Golder and Huberman [5] analyse data gathered from a popular folksonomy, del.icio.us, to better understand the structure of tagging systems, as well as their dynamic aspects. The study examines how tags are used over time by users and how they eventually stabilize, for a specific resource, over time. Findings show a high variety in the sets of tags employed by users. The study also discovers some measure of regularity in user activity, tag frequencies, kinds of tags used and bursts of popularity in bookmarking. On this basis, a dynamic model is proposed to predict stable tagging patterns. Golder and Huberman also discuss difficulties of folksonomies related to semantic and cognitive aspects of tags, such as problems in synonymy, polysemy and basic level variation.

To overcome common issues in using tags, some new approaches, bent to exploit semantic techniques in collaborative tagging, have recently appeared in literature.

Heymann and Garcia-Molina [9] describe an algorithm which aims to address the basic level problem by converting a large corpus of tags into a navigable hierarchical taxonomy. Tags are aggregated into vectors denoting the number of times a tag has been used for each annotated resource. A similarity function is defined and calculated using the cosine similarity between vectors, then a threshold is established to prune irrelevant values. Finally, for a given dataset a tag similarity graph is created exploiting the social network notion of graph centrality. Starting from the similarity graph and according to three fundamental hypotheses, namely hierarchy representation, noise and general-general assumptions, a latent hierarchical taxonomy is mined. The algorithm is tested on two different datasets, from del.icio.us and CiteULike, but only in the first case results are promising.

In [17] the authors propose statistical techniques to mine the implicit semantics embedded in the different frequencies of co-occurrences among users, resources and tags in folksonomies. A probabilistic generative model is used to represent the user's annotation behaviour in del.icio.us and to automatically derive the emergent semantics of the tags. Synonymous tags are grouped together and polysemous tags are identified and separated. Moreover, the derived emergent semantics is exploited to discover and search shared web bookmarks. The initial evaluation shows that such a method can effectively discover semantically related bookmarks.

Similarly to the previous work but with different aims, in [1] clustering techniques are applied to folksonomies. The authors propose a more effective searching, subscribing and exploring service which automatically generate clusters of tags. For tags used for the same resource, the algorithm counts the number of co-occurrences of any pair of tags. A cut-off point is established to distinguish between strongly and weakly related tags which are represented in an undirected weighted graph. The authors exploit an algorithm for graph clustering which introduces the modularity function for measuring the quality of a particular partition of nodes in a graph. The algorithm is tested on the RawSugar database and results are available at the RawSugar lab page¹.

Xu et al. [18] define a set of general criteria for a good tagging system to identify the most appropriate tags, while eliminating noise and spam. These criteria, identified through a study of tag usage by real users in My Web 2.0, cover desirable properties of a good tagging system, including high coverage of multiple facets to ensure good

¹ <http://www.rawsugar.com/lab>

recall, least effort to reduce the cost involved in browsing, and high popularity to ensure tag quality. The authors propose a collaborative tag suggestion algorithm using previous criteria to recommend high-quality tags. The proposed algorithm employs a goodness measure for tags derived from collective user authorities to contrast spam. The goodness measure is iteratively adjusted by a reward-penalty algorithm, which incorporates content-based and auto-generated tags. These principles ensure that the suggested tag combination has a good balance between coverage and popularity. Preliminary results, coming from an experiment conducted on My Web 2.0, show that such an algorithm is effective in suggesting appropriate tags that match the expected properties.

Schmitz [14] explores a model that leverages statistical natural language processing techniques to induce ontology from Flickr database. An existing probabilistic subsumption based model is adapted to existing tags set, adjusting the statistical thresholds to reflect the ad hoc usage, and adding filters to control the highly idiosyncratic Flickr vocabulary. The proposed model produces subtrees that generally reflect distinct facets, but can not categorize concepts into facets. Although resulting trees are manually evaluated, early results are promising compared to related subsumption models. Moreover a series of refinements to the model are planned to improve the accuracy and induce a faceted ontology.

Finally, Gruber's proposal [6] is the establishing of a common format to achieve the interoperability among different tagging applications. Tags are considered as a form of voting and the act of tagging performed by users is compared to the innovation of incorporating the hyperlink as a measure of popular acclaim pioneered by Google. Two use cases, illustrating current issues in sharing tags across multiple applications, are presented as motivations of the need of a common conceptualization for representing tags meaning. The author proposes to create an ontology for tagging, the "TagOntology", meant to identify and formalize a shared conceptualization of the tagging activity, as well as to develop the technology that commits to the ontology at the semantic level.

5 Towards a Collaborative Approach for Ontology Evolution

While most efforts, such as the above mentioned works, are focused on how to create better folksonomies, we propose an approach going to the reverse direction. Our goal is to exploit what is already provided by existing ontologies, such as explicit semantics and support for reasoning in combination with the ability of folksonomies to foster collaboration within a community.

The vision we propose is a community of autonomous and networked users who cooperate in a dynamic and open environment. Each participant will organize some piece of knowledge according to a self-established vocabulary, create connections and negotiate meaning with other users within the community.

Augmenting the involvement of users, by enabling community members to actively participate to the ontology evolution process is a key factor to achieve a community common ground. Starting from an existing ontology and allowing users to freely edit ontology classes, according to their personal vocabulary, can significantly improve the ontology maintenance process, complying with the knowledge drift.

However, current ontologies are usually manually written in standard formal languages, such as OWL², through standalone toolkits, such as Protégé³. The requirement of background skills, such as the knowledge of the OWL syntax, and the lack of tools enabling a distributed and collaborative contribution to the ontology enrichment can severely hinder our vision.

Given an initial ontology, we propose a collaborative ontology evolution system, which allows community members to add, modify, or delete existing and new ontology classes, according to their own needs. The editing of the classes does not require any special skill but it is allowed through the use of simple metadata, like adding tags in current tagging systems.

An open rating system will also be provided, each time a user contributes to the evolution of the ontology this can be seen as a voting activity. Adding or editing a class can be seen as a vote for the new class added, while contributing to the ontology without editing existing classes (e.g. inserting new individuals) can be considered as a vote for an existing class. A measure of the quality of an ontology class can be thus calculated according to the weight average of all the votes obtained by the class. This rating system can lead the most popular names for a specific ontology class to rise to the top, while the less exploited ones will fall to the bottom, similarly to the behaviour of a tag cloud in a folksonomy.

On the other hand, to support this collaborative ontology evolution, an environment providing distributed access and supporting itself group activity, such as, simultaneous viewing and editing, is needed. While distributed access can be provided by a common web interface the collaborative work can be ensured by the wiki technology.

Wikis are themselves collaborative tools providing a community with web writing and browsing functionalities. Wiki systems typically provide an easy-to-use editing environment to create or modify content on the fly requiring no tool but the browser. They usually have a version control system to record modification of contents in order to show-up recent changes and the versions history, a user profile and concurrent conflict management system to enable multiple user editing the same contents, and a content navigation system that simplifies indexing, searching and linking wiki pages within a wiki system.

On this basis, our purpose is to use the wiki technology to edit ontologies via the web, thus developing a Web Ontology Editor. A web-based interface will provide features to support collaborative editing, ontology evolution tracking (e.g., identify classes that have been added, deleted or modified), as well as browsing functionalities that allow users to search and navigate the ontology.

An ontology module can be composed of one or more wiki pages, multiple users can edit the same content with version control and transaction management, and the ontology can be managed like a common wiki repository. Moreover, most wikis offer the opportunity to define customized markup. Users can associate predefined sequences of characters with commands that the wiki engine can interpret and execute, in order to render a new sequence of characters expressed in a different syntax. Therefore it is possible to define a set of custom markup tags corresponding to

² <http://www.w3.org/TR/owl-features/>

³ <http://protege.stanford.edu/>

the syntax of the ontologies, such as OWL or RDF syntax. When a wiki page is under editing, its custom markup is translated to user friendly text, such as HTML web page. This customized Wiki markup can be expressed in human readable syntax and like in current wikis users could create new contents simply by adding and modifying the source of an existing one.

Adopting a collaborative approach for ontology maintenance is a challenging research topic for the benefits it can bring to conventional approaches [3].

Ontologies which are improved and used as a community reflect the knowledge of users more effectively than ontologies maintained by knowledge engineers who struggle to capture all the variety taking place within a lively community.

Collaborative ontology editors can strengthen community participation in ontology development and maintenance process because users are enabled to autonomously change knowledge and look at changes that are triggered by their actions.

Furthermore, classic ontology maintenance is expensive as one or more knowledge engineers have to be called on purpose. With community-driven ontology evolution, costs are split over a wide group of people who have a special interest in maintaining the ontology they use up-to-date.

The proposed approach can be a first step toward a collaborative system capable of allowing ontologies to evolve mainly through the contribution of its users. A web ontology editor can relieve users of the system to know OWL syntax and allow them to contribute to the ontology by adding tags as proxies of metadata.

Future work plan includes the development and evaluation of the proposed web ontology editor. We aim to show that such an environment can scale up and support collaboration among several users. A further step will be to explore how semantics can be achieved through collaboration among users without burdening the user experience.

References

1. Begelman, G., Keller, P., Smadja, F.: Automated Tag Clustering: Improving search and exploration in the tag space. Proc. Of WWW2006, Collaborative Web Tagging Workshop (2006)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (2001)
3. Davies, J., Fensel, D., van Harmelen, F.: Towards the Semantic Web: Ontology-driven Knowledge Management. John Wiley & Sons, Inc., New York, NY (2003)
4. Furnas, G. W., Landauer, T. K., Gomez, L. M., Dumais, S. T.: The vocabulary problem in human-system communication. Communications of the ACM 30, 11 (1987) 964-971
5. Golder, S., Huberman, B.: Usage patterns of collaborative tagging systems. Journal of Information Science 32, 2 (2006) 198-208
6. Gruber., T.: Folksonomy of Ontology: A Mash-up of Apples and Oranges. First on-Line conference on Metadata and Semantics Research (MTSR'05) (2005)
7. Gruber., T.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies 43 (1993) 907-928
8. Haase, P., Volker, J., Sure, Y.: Management of dynamic knowledge. Journal of Knowledge Management 9, 5 (2005) 97-107

9. Heymann, P., Garcia-Molina, H.: Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report InfoLab 2006-10, Department of Computer Science, Stanford University, Stanford, CA, USA (2006)
10. Mathes, A.: Folksonomies-Cooperative Classification and Communication Through Shared Metadata. Technical Report LIS590CMC, Computer Mediated Communication, Graduate School of Library and Information Science, University of Illinois (2004)
11. McCulloch, E., Macgregor, G.: Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool. *Library Review* 55, 5 (2006)
12. Merholz, P.: Clay Shirky's Viewpoints are Overrated. (2005)
13. O'Reilly, T.: What is Web 2.0. (2005)
14. Schmitz, P.: Inducing Ontology from Flickr Tags. Proc. Of WWW2006, Collaborative Web Tagging Workshop (2006)
15. Shirky, C.: Ontology is Overrated: Categories, Links, and Tags. (2005)
16. Vander Wal, T.: Folksonomy Definition and Wikipedia. (2005)
17. Wu, X., Zhang, L., Yu, Y.: Exploring social annotations for the semantic web. Proc. of the 15th international conference on World Wide Web (2006) 417-426
18. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the Semantic Web: Collaborative Tag Suggestions. Proc. Of WWW2006, Collaborative Web Tagging Workshop (2006)

Knowledge Sharing over Social Networking Systems: Architecture, Usage Patterns and Their Application

Tanguy Coenen, Dirk Kenis, Céline Van Damme, and Eiblin Matthys

Vakgroep MOSI, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
{tanguy.coenen, dirk.kenis,
celine.van.damme, eiblin.matthys}@vub.ac.be

Abstract. The recent success of social networking sites like mySpace, Friendster, Orkut, LinkedIn, Ecademy and openBC indicates that extending one's social network through a virtual medium is a popular proposition. The social networking paradigm can be integrated with the knowledge management field, in which sharing knowledge with others is a central issue. This paper investigates how the social networking concept can be applied to support knowledge sharing between people. Together with common features in the architecture of social networking systems, a number of platform independent usage patterns are discussed that can support knowledge sharing between members. Finally, we present the open source KnoSoS system, which integrates the discussed architecture and usage patterns.

Keywords: social networking, knowledge management.

1 Social Networking Systems

New computer mediated technologies, like instant messaging, voice-over-ip and video telephony have significantly lowered the cost of communication and, when compared to email, have made computer-mediated communication much richer. It is now possible to talk hours at a stretch to someone at the other side of the world at no cost beyond the basic Internet connection fee. Yet these technologies still do not provide all the cues that are available in a face-to-face conversation. When meeting a person over the Internet and interacting with him, it is difficult to find out if you are really interacting with who the person claims to be. Furthermore, finding people on the Internet with whom you are likely to have interesting and useful conversations is not something that is likely to occur by chance. To make sure that people can represent their identity and to allow people with mutual or compatible interests to find each other, social networking systems have been created. The first social networking systems were dating systems, to which people turned to find a partner. Yet dating systems are not the subject of this paper, as they offer little to support knowledge sharing.

Through systems like Orkut, MySpace, LinkedIn, O'Reilly Connection, Ecademy and OpenBC, the social networking paradigm is spreading quickly on the Internet as a way for people to develop an online social life. Spearheading this change are American social networking systems like Friendster, Orkut and MySpace. In May 2006, the latter had 76.524.752 members, which is roughly equivalent to the population of

countries like France or Germany. On 19/06/2005, mySpace was sold to Rupert Murdoch's News Corp for 580 million \$, indicating that social networking systems are very actual material and that, if they are not yet the business of today, they are probably the business of the future.

Considering the success of systems like MySpace, it is obvious that young people are currently spending a lot of their time online, using computer mediated communication and social networking systems to express their personality and meet with friends. It is likely that, as these people grow older and enter the work force, they will be accustomed to these technologies and will expect to be able to use them in their work environment.

By presenting the common architecture of social networking systems and a number of general usage patterns, this paper indicates how social networking systems can support knowledge sharing. This is relevant both within large organisations, between organisations or between individuals without organisational affiliation. Yet before indicating how social networking systems can benefit knowledge sharing, it is necessary to present an overview of some issues related to the sharing of knowledge.

2 Knowledge Sharing

One of the central issues in the knowledge management field has always been the sharing of knowledge [12]. Knowledge sharing can occur in what we call the passive and the interactive mode. In the passive mode, the source, who owns the knowledge, externalises his knowledge and stores it as information. The receiver, who wishes to use the knowledge, assimilates the knowledge but has no way of formulating feedback to the source. Unlike what is the case for passive knowledge sharing, interactive knowledge sharing involves a possibility for the receiver to provide the source with feedback. From a constructivist perspective, individuals are seen as possessing their own unique understanding of the world. As a consequence, communication is by definition complicated as it confronts different mental models. [7]. The possibility to produce feedback can thus be essential in situations where the receiver does not understand the information, provided by the source. The source can then re-formulate his knowledge in a way that is more suited to the needs of the receiver¹

Passive knowledge sharing has the great benefit of being highly reusable. The source externalizes his knowledge once and the resulting information can be reused many times by different receivers. Yet the knowledge, which is made available through passive knowledge sharing, can go quickly out of date and there is a motivational problem. Indeed, people find it hard to contribute their knowledge to a vague audience if they do not have a clear view on the "return on investment" which they will obtain from sharing their knowledge [11].

Therefore, knowledge sharing must not only focus on passive knowledge sharing, but should also support interactive knowledge sharing. Both modes are useful and should therefore be present in and between organisations or between independent

¹ Also, complex knowledge, which is highly intertwined with other knowledge components, requires an interactive mode of knowledge sharing, as the receiver may need to obtain some context knowledge in order to correctly understand the knowledge of the source.

individuals. Support for the passive mode has been around for over a decade in the form of information management approaches, focussing on storing information in databases. We argue that a focus on the interactive mode of knowledge sharing is necessary and that applying concepts of social networking systems in addition to new rich computer-mediated communication technologies can support this. In the next section, the common architectural features of social networking systems are presented.

3 Common Features of Social Networking Systems

A study of existing social networking systems (mySpace, Friendster, Orkut, LinkedIn, Ecademy and openBC) has revealed a common architecture, depicted in figure 1 [2][3]. In the individual space, the user is allowed to create a personal profile, containing structured and unstructured information. The structured information part of the profile contains information on different facets of the individual's personality. These facets vary between systems. Indeed, where some systems concentrate more on creating friendship relationships (e.g. mySpace, Friendster and Orkut), others are focused on creating business relationships (e.g. LinkedIn, Ecademy and openBC). Information on for example one's musical taste would be interesting to have in a friendship-oriented system, but less interesting in a business-oriented system. In practice, each social networking system contains different fields in which structured information can be entered. The unstructured fields allow people to create a free and rich representation of their own identity. Wysiwyg editors are provided to create webpages containing text, images, movies and sound clips. In addition, some systems allow people to create blog entries, which further enrich a person's profile.

In the dyadic² space, users create contacts with whom communication can be undertaken over the internal messaging system. This internal messaging system is very similar to an email system, but has the advantage of being able to shield a user's external email address. The message, which is sent over the internal messaging system, is forwarded to the user's external email account, without the need for the sender of the contact to know the email address of the receiver. This is a necessary measure to prevent social networking systems from becoming vehicles for email spamming.

Another element of the dyadic space is the possibility for both members of a dyad to create feedback on the other member of the dyad. In different analysed social networking systems, this can be done by rating certain characteristics of the other member of the dyad, or by writing testimonials on this person [2]. We call such signals, "identity feedback" and have found it to be important in the creation of a sense of trust in social networking systems. Indeed, such systems create many relationships that are purely virtual in nature, which results in a need for the members of the system to evaluate the genuineness of the system's users.

Finally, the group space contains tools, which allow knowledge sharing between multiple people. This space constitutes the overlap between the areas of knowledge management, social networking systems and community informatics. In most systems, the tools, which are available in the group space, are limited to a forum on

² A dyad is a relationship which has been acknowledged by both members of a relationship.

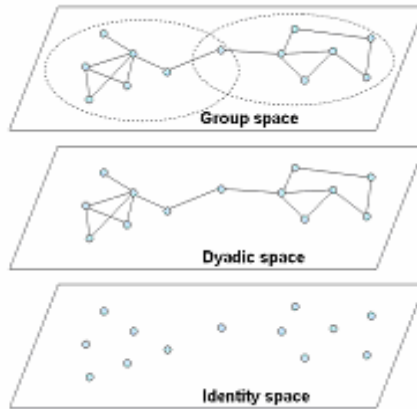


Fig. 1. General structure of social networking systems

which members can post and read messages. There is still much room for improvement in the group space, as it exists in current social networking systems, in order to better support knowledge sharing between multiple people.

4 Pattern Languages

Patterns describe common ways of solving problems. A pattern language is an ordered set of patterns that are tied together in a logical way. Problems arise in all domains of human activity, making pattern languages a useful way to structure problems and solutions in different disciplines. Whereas the term was first introduced in architecture [1], pattern languages have been applied to many other fields, like education and object-oriented programming. In this paper, we do not claim to develop a pattern language, but present a number of loose patterns, which are the result of our analysis, development and usage of social networking systems for knowledge sharing.

Each pattern has a name, a context, a system of forces and a solution. The context describes the conditions that must be taken into account to fully understand the problem and its solution. The system of forces describes the problem at hand and the solution describes how the problem is solved. An important point concerning patterns is that they are meant to be generic, applying to many different instances of the same problem. Translated to social networking systems for knowledge sharing, this means that we aim to formulate patterns that are applicable to social networking system in general, and should therefore be taken into account when designing such systems. The added value of a pattern does however not only lie in the solving of a problem. Indeed, many others, like [13], see messy situations, which are the characteristic of human activity, as also problematic in their definition of the *problem*, besides the solution. In other words, identifying the right problem is a challenge in itself when dealing with human activity.

The patterns which are presented next are the result of our experience with social networking systems (mySpace, Friendster, Orkut, LinkedIn, Ecademy and openBC)

and our insight in the issue of knowledge sharing. Other patterns have been developed or are in development (e.g. identity representation, rating of content, project management), but cannot be discussed here due to a lack of space.

4.1 Pattern 1: Creating Group Boundaries

*Context*³

Previous research [2][3] has established that there is a statistically significant difference in the amount of exchanged messages between open and closed groups. The difference lies in the boundaries that exist around these groups. In an open group, everyone can participate in the communications that are a part of the group's knowledge sharing activity. In a closed group, one needs to be a member⁴ to participate in the group's knowledge sharing activity. In the 1196 open groups of the Ecademy social networking system, 3 times less messages were exchanged on average, compared to the 304 closed groups in the system.

This is in line with [9] who proposes that knowledge sharing will be more common in groups with a stable membership, due to a lower expectation of free-rider behaviour by other members of the group. Another point, which can support this higher knowledge sharing activity in closed groups, is the argument in [2] that the value of knowledge lies in the scarcity of the capacity to act that it produces. If one gives away one's knowledge to the whole world, as is the case in open groups, the scarcity of this capacity to act can seriously decrease, as others can readily assimilate the knowledge. Yet contributing knowledge to a group of people who share a certain interest can produce reciprocation benefits, originating from generalized social exchange between the members of the group [9]. If this group is closed, the contributor will feel he may receive these benefits, while only sharing his knowledge with a small subset of humanity. Thus, the benefits in the case of knowledge sharing in a closed group are more likely to outweigh the costs, due to the targeted nature of the knowledge-sharing act.

Still, not all knowledge is essential to the economic position of the holder of the knowledge. Thus, some knowledge can be easily donated without harming the position of the source. If this is the case, it can be in the interest of all the participants of the system and not only the members of a certain group to be able to access the posted content. In this case, posting to an open group seems opportune.

In sum, both open and closed groups are necessary in social networking systems for knowledge sharing, making the creation of boundaries a central pattern in their functioning. The context, described in this paragraph, allows us to formulate 2 patterns: creating groups with or without boundaries and creating content in groups. These patterns are discussed next.

Problem

How to create groups with and without boundaries?

Solution

Every user should be able to set up a group if he wishes to do so. At the time of creation, the person who takes the initiative for the creation of the group decides if the group should have an open or closed membership. In a closed group, mechanisms

³ The context, described in this section, is relevant to the first two patterns which are presented.

⁴ Membership is often obtained by requesting it explicitly to one or more administrators.

should be provided which allow outsiders to apply for group membership. One or more group administrators should then be able to screen the application and grant or deny access to the group.

4.2 Pattern 2: Tracking Content

Context

A social networking system, like the Internet in its totality, is an environment in which content is produced at a very high rate. The lowering of the content production threshold is what probably best defines the web2.0 paradigm. Indeed, systems that allow users to easily create and publish text, movies; audio clips etc are producing an explosion of content. This has accelerated the rate at which content is produced, but has also lead to an increased heterogeneity in content. This combination of high production rate and content heterogeneity has resulted in a situation in which the structuring of content by means of hierarchies or ontologies has become unfeasible.

Problem

How to keep track of heterogeneous content, which is produce at a high pace?

Solution

Allow users to add freely chosen tags to describe the content in the system in order to retrieve it later. In doing so, they create metadata, expressed in their own terms. A consolidation of all these tags constitutes a bottom-up taxonomy, created by the members of the social networking system. This taxonomy is a reflection of the vocabulary, prevalent in the minds of the users. The absence of a controlled vocabulary and the immediate return on investment of retrieving tagged content by using one's own words, is a real incentive for tagging content and tagging it a conscientious way.

4.3 Pattern 3: Grasping Perspectives

Context

During a knowledge sharing process, it is important to be able to grasp the perspective of an individual or a group of individuals [5]. By providing users with a mechanism for creating boundary objects, it is possible to create a perspective that can then be used by others to quickly find out what the cognitive schemes of other people look like. Indeed, a boundary object by definition is an object that can be shared across perspectives [4]. This can facilitate knowledge sharing, as, according to social science perspectives like constructivism [7], different people hold different views of what constitutes reality. Therefore, visualising the perspectives of people through boundary objects can increase the insight that different parties of a communication have in each other's perspectives. This allows them to conduct a communication that is more targeted to the perspective of the other participants in the communication.

Problem

How to create a visual representation of the cognitive perspective of a person or of a group of persons?

Solution

Use the tags which where attributed as a result of the previous pattern (content tracking) to create a cognitive network view which represents the way different tags are

related. This can be done, based on the principle of co-occurrence [8], by which two tags that reference a same resource, accessible through a URL, are linked by a relationship of strength 1. The strength of the relationship between the tags is incremented by 1 each time the 2 tags are used together to reference a concept. Thus, if knowledgeSharing and socialNetworking have been used together 20 times together to reference a resource, their association’s strengths will be 20. This approach can be used to create visualisations of independent perspectives, but can also be used to merge individual perspectives into a group perspective.

5 Knosos

In this section, we present the open-source KnoSoS system and the way in which it implements the architecture and usage patterns which have been developed earlier. As the KnoSoS system is open-source and freely available, it is possible to use the system for knowledge sharing and to quickly develop and test new additions to the social networking paradigm. Hence it can be deployed in and customized to different environments, like multinationals, governmental or non-governmental organisations. Furthermore, KnoSoS implements the architecture and all the patterns which have been discussed., which is not the case for all social networking systems (cf. in tables 1 and 2).

Table 1. Support of the 3-layer architecture by different social networking systems

System	Individual space	Dyadic space	Group space
OpenBC	X	X	X
Ecademy	X	X	X
Orkut	X	X	X
mySpace	X	X	X
O’Reilly Connection	X	X	-
linkedIn	X	X	-
KnoSoS	X	X	X

Table 2. Support of the 3 patterns by different social networking systems

System	Pattern 1	Pattern 2	Pattern 3
OpenBC	X	-	-
Ecademy	X	-	-
Orkut	X	-	-
mySpace	X	-	-
O’Reilly Connection	-	-	-
linkedIn	-	-	-
KnoSoS	X	X	X

Table 1 shows that most social networking systems support the 3-layered architecture. In table 2, however, it is clear that of the proposed patterns, only pattern 1 seems to be supported by most social networking systems. Patterns 2 and 3 are only supported by KnoSoS. Yet these patterns are important to the support of knowledge sharing. Therefore, KnoSoS is probably more suited for the support of knowledge sharing than are other existing systems.

Subsequently, the implementation of the 3 usage patterns in the KnoSoS system is addressed briefly. For a detailed description of KnoSoS and how this system implements the 3-layer architecture as well as present and other patterns, the reader can visit and test the system⁵.

5.1 Pattern 1: Creating Group Boundaries

KnoSoS, like a number of other social networking systems, allows the administrator of the group to make it open or closed. If a user wants to access an open group, this can be done without extra effort. To become member of a closed group, however, the user needs the approval of the administrator.

5.2 Pattern 2: Tracking Content

The tracking of content is done by allowing users to add tags to the different types of content which are available in KnoSoS. The basic content types are blogs, forum posts and books. The latter contain collaborative writing features. Additional custom content types can be added by the administrator of the system. This is useful for creating repositories of hyperlinks, papers, idea proposals, etc... Each of these content types can be tagged by the user who visits the specific content. In addition to allowing the user to keep track of the content in the system, the tagging functionality provides metadata which will allow subsequent development of more advanced features, like the matching of content to users, the matching of users to each other and the development of bottom-up ontologies. Whereas these topics require further research, one use of the metadata created by users has already been implemented in the form of boundary objects. This is discussed subsequently.

5.3 Pattern 3: Grasping Perspective

The tags, produced through the application of pattern 2, are used to create a map of the perspective of a user. This is done by means of a Java applet which we developed, called TagViz⁶. The relationships between the tags are inferred based on the co-occurrence method, described in section 4.3.

At the time of writing, only the perspectives of single users could be visualised, but efforts are under way to visualise the perspectives of sets of users. In this way, it will be possible to visualise the perspective of a whole group. This visualisation of

⁵ The system can be tested at www.knosos.be. KnoSoS is a distribution of the open-source Drupal content management systems to which custom modules have been added.

⁶ TagViz is available for testing at <http://www.knosos.be/sandbox/tagviz/index.html>

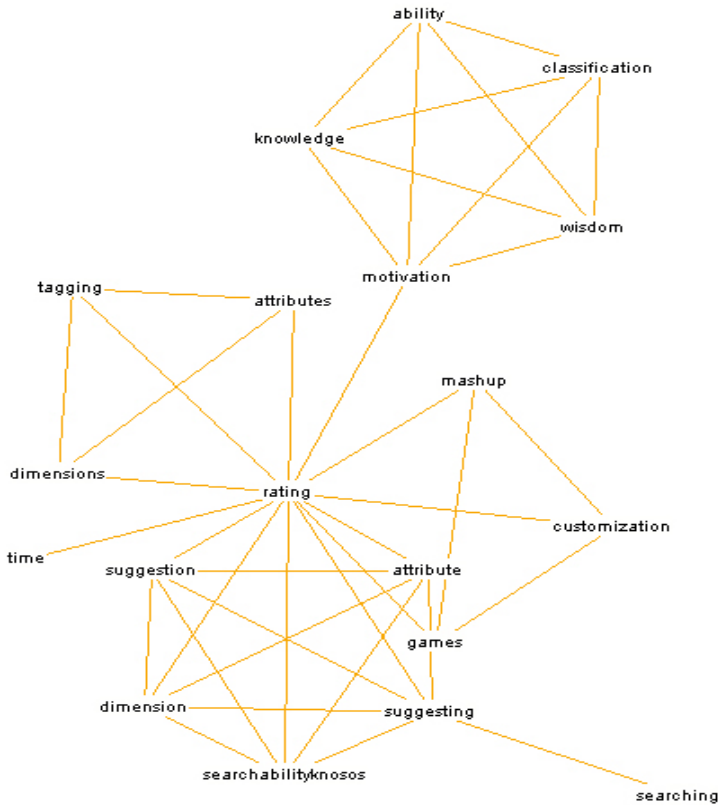


Fig. 2. Visualisation of part of a personal perspective through a boundary object, constructed from the tags in a KnoSoS account

perspectives is expected to facilitate knowledge sharing between users and the localisation of users with specific knowledge. Whether this is truly the case offers an interesting opportunity for future research.

6 Conclusion

This paper proposes that passive knowledge sharing should be complemented with interactive knowledge sharing approaches and that the social networking paradigm is well suited to allow this. From experiences by participating in -and designing- social networking systems, 3 patterns have been distilled to enhance social networking systems for knowledge sharing. Furthermore, it was discussed that most present social networking systems share a common architecture based on 3-layers. .

The open-source KnoSoS system builds on this common architecture and implements the presented patterns. At the moment it acts as a platform in which new techniques can be tested. One particular improvement which will be pursued in the

immediate future is the matching of users to each other and the matching of users to content. The tags which are produced as a consequence of using the systems will play a major role in this. A beta version of the KnoSoS system can be tested on www.knosos.be.

References

1. Alexander, 1977, *A Pattern Language: Towns, Buildings, Construction*, New York: Oxford university press
2. Coenen, T, 2006, *Knowledge sharing over social networking systems*, forthcoming PhD dissertation
3. Coenen, T, 2006, *Structural aspects of social networking systems*, Proceedings of Web-based communities 2006, San Sebastian, Spain
4. Carlile, P R, 2002, *A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development*, *Organisation Science*, Vol 13, No 4, p 442-455
5. Boland, R J, Tenkasi, R V, 1995. *Perspective making and perspective taking in communities of knowing*. *Organisation Science*, Vol 6, p350-372.
6. Daft, R, Lengel, R, 1986, *Organisational Information requirements, media richness and structural design*, *Management science*, Vol 32, p554-571
7. Fosnot, C.T., 1996, *Constructivism: a psychological theory of learning*, in Fosnot, C.T., 1996, *Constructivism – theory, perspectives and practice*, , New York: Teachers Press College
8. Heylighen, F, 2001b, *Mining Associative Meanings from the Web: from word disambiguation to the global brain* in Temmerman, R (ed.), *Proceedings of Trends in Special Language and Language Technology*, Brussels: Standaard Publishers, <http://pespmc1.vub.ac.be/Papers/MiningMeaning.pdf>
9. Kollock, P, 1999, *The economies of online cooperation*, in Smith, M A, Kollock, P (eds), *Communities in cyberspace*, London: Routledge, p220-239
10. Markus, M L, 2001, *Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success*, *Journal of Management Information Systems*, Vol 18, No 1, p57-93
11. Nahapiet, J, Goshal, S, 1998, *Social capital, intellectual capital and the organisation advantage*, *Academy of Management review*, No 23, p242-266
12. Nonaka, I, Takeuchi, H, 1995, *The knowledge-creating company*, Oxford University press, New York
13. Vennix, J, 1996, *Group model building: facilitating team learning using system dynamics*, Chichester: Wiley

Metadata Mechanisms: From Ontology to Folksonomy ... and Back

Stijn Christiaens

Semantics Technology and Applications Research Laboratory
Vrije Universiteit Brussel
stijn.christiaens@vub.ac.be

Abstract. In this paper we give a brief overview of different metadata mechanisms (like ontologies and folksonomies) and how they relate to each other. We identify major strengths and weaknesses of these mechanisms. We claim that these mechanisms can be classified from restricted (e.g., ontology) to free (e.g., free text tagging). In our view, these mechanisms should not be used in isolation, but rather as complementary solutions, in a continuous process wherein the strong points of one increase the semantic depth of the other. We give an overview of early active research already going on in this direction and propose that methodologies to support this process be developed. We demonstrate a possible approach, in which we mix tagging, taxonomy and ontology.

Keywords: tagging, folksonomy, community informatics, faceted classification, ontology, Semantic Web.

1 Introduction

The vast amount of information currently available can make finding the exact piece of information a tedious process. Despite today's raw search engine power, finding what you seek sometimes requires something more fine-grained, something more specific. For instance, you visit your town's website when you try to find a form for their administration, not a generic search engine. The problem in today's information sources is that the data is highly unstructured [1] or semi-structured and that meaning is only visible to human agents. Homepages, blogs, forums and others contain valuable community-produced knowledge, but the search engines have difficulties identifying and retrieving the knowledge you are looking for.

The Semantic Web [2] is the next generation of the WWW, a Web in which all content has machine-processable meaning. This Semantic Web provides all the functionality needed to build the Pragmatic Web [3,4] on top of it. Communities will no longer search, but rather find and use information in this Pragmatic Web. The explicit meaning, understandable by both human and machine agents, attached to content is necessary for proper information retrieval and usage.

It is clear that mechanisms are needed to incorporate or annotate content with semantics, with some form of meaning in order to increase machine-understanding. Research on the Semantic Web as well as current trends in the Web (the so-called Web 2.0 [5])

resulted in several mechanisms. However these mechanisms do not all provide the same semantic depth. They have different goals, users and granularities.

In section 2 we give an overview of currently existing mechanisms. We describe their purpose and their main strong and weak points. We then discuss how these mechanisms could benefit from the others in section 2.3 and list some early research in this direction. We give an example of mechanism cooperation and how it benefits communities in section 3. Finally, we end with conclusions and suggestions for possible future work.

2 Meaning Mechanisms

In this section we give a brief overview of several meaning mechanisms. It is not our intention to fully describe each of these mechanisms, but to introduce (or refresh) them to the reader for facilitation of further sections.

2.1 Overview

The earliest form of metadata to describe meaning is the introduction of keywords. These are labels with which the author (creator, publisher, ...) of the content describes his content. In the early days of the web, these keywords could (and can still) be used by search engines for information retrieval.

Tagging is the process of describing the *aboutness* of an object using a tag (a descriptive label). In the current (so-called) Web 2.0 trend (e.g., <http://del.icio.us>), tagging is done by the observers of the content. These tags can now be shared and used by all members of the community. The organic organization that grows through this sharing was coined folksonomy by Thomas Vander Wal [6]. Vander Wal [7] distinguishes between a broad and a narrow folksonomy. In a broad folksonomy all users can tag the visible content, while in a narrow folksonomy only the author tags his content.

A taxonomy is a hierarchical classification of things. It is mostly created by the designer of the system or a knowledge engineer with domain knowledge. Authors (or categorizers) must find a good place in this hierarchy to position their content.

In faceted classification [8], objects are described by facets. A facet is one isolated perspective on the object (e.g., color of wine). Each facet has a set of terms that are allowed (e.g., red and white for wine color). According to Kwasnick [9] there are many advantages in faceted classification, in particular the flexibility and the pragmatic appeal.

Gruber [10] writes that an ontology is *an explicit specification of a conceptualization*. An improved definition was provided by Borst [11]: *ontologies are defined as a formal specification of a shared conceptualization*. Ontologies are seen as the technology to enable the Semantic Web and many ontology languages and approaches have been developed [12], for instance RDF [13] and OWL [14].

2.2 Comparison

We will focus on the main contributions of the meaning mechanisms described in the previous subsection. It is not our intention to perform a full-scale analysis of all mechanisms, but rather to identify major strengths and weaknesses.

The author-created keyword provides a very precise (high quality) view on the content. Unfortunately, they form a one-person perspective. His information can never equal the amount that an entire community can deliver (low quantity). Narrow folksonomies suffer the same disadvantages as author-created keywords as they are very similar¹. Broad folksonomies break free of the one-person-perspective and deliver an entire community-view on the content. The drawback is of course the quality of the metadata. For instance, a tag "toread" is not very useful except for the person who labeled the content that way. A folksonomy can also be used to compile a tag profile for users. This way, both content and people can be found (see e.g. [15]).

On the other hand, we have the more heavy-weight mechanisms. High quality taxonomies, facet classification and ontologies are costly to create and maintain. The most expensive in creation and maintenance is an ontology. It requires consensual agreement on its contents from community members. Their main benefit is that these mechanisms deliver very rich meaning. For instance, if information is annotated by means of an ontology, it can be queried for specific information using a conceptual query language (e.g., [16]), much like a database can be queried (e.g., using SQL²). However, given that most searches on the web are restricted to one or two keywords, it seems not likely that people will learn to use such a query language. These mechanisms seem the most distant from real-world users and are criticized as such (e.g., [17]).

2.3 Mechanism Collaboration

As we explained in the subsection, each mechanism has its own advantages when compared to the others. We propose to divide the mechanisms in two groups: free and restricted. A free mechanism is one that allows anyone (both creator and observer) to annotate the content with any label he desires. A restricted mechanism is one that fixes the metadata, and all content must satisfy it. Observers can not annotate the content with their own labels. Free mechanisms are popular, but receive critique that the resulting information might not be of the desired quality. Restricted mechanisms seem less popular, but provide higher quality metadata. However, they are said to be too static, too inflexible for the ever evolving real-world situation.

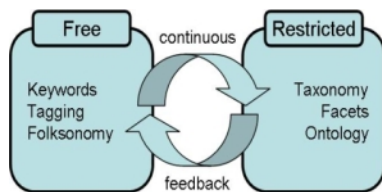


Fig. 1. Continuous feedback loop

¹ Similar, but not the same, as keywords are placed in technology (e.g., HTML) and for use by agents (e.g., a crawler) and narrow folksonomies are displayed next to the content for use by both human and machine agents.

² <http://en.wikipedia.org/wiki/SQL>

In Figure 1 the listed mechanisms³ are positioned according to our division. The feedback loop in the figure indicates *the way to go* in meaning mechanisms. This closely corresponds to the SECI model [18], where tacit knowledge is made explicit, combined, and converted again to add to the existing tacit knowledge. Ideally, a third zone should be created, a gray area where meaning can benefit from both free and restricted mechanisms⁴. In our opinion, this will result in a system that receives the benefit from the free zone (quantity) as well as from the restricted zone (quality).

Research in this gray zone is in its early stages and seems mainly focused on pushing systems in the free zone towards more quality (e.g., [19,20]). We believe this is because of the current popularity of these systems. Other research (e.g., [21,22]) takes meaning in the restricted zone as the seed in order to guide users to richer annotation. In the ontology world, we mention DOGMA-MESS [23], an ontology engineering methodology for communities. It succeeds in bringing the restrictedness of ontology a step closer to the free zone in a *messy*, but structured and guided process.

Apart from improvements based on benefits of one side (either free or restricted), methodologies and mechanisms should be created that actually use and reinforce the gray zone (and thus both sides as well).

3 The Guide: A Research Lab's Memory

3.1 Folksonomy and Taxonomy

The Guide is the community portal in STARLab⁵. The community members are the people working at STARLab. The Guide is powered by Drupal⁶, an open source content management system. Its modular approach and multitude of possibilities deliver an endless plethora of possibilities for *community plumbing*. At STARLab, we use the Guide to remember and retrieve all useful information (ranging from technical issues over team building to deep theoretical discussions). Some content is completely free (e.g., a blog post can be about anything), while other information has to be placed in special containers (e.g., book pages and forum topics). All posts have to be tagged by the author herself. This way a narrow folksonomy emerges in the Guide, which we can visualize in a tag cloud⁷.

Figure 2 displays the Guides' tag cloud at the age of month six⁸. The figure also displays part of a taxonomy we distilled from the tag cloud. We analyzed all available tags in order to get a good insight in the emerging categorization. We then grouped

³ Note that our list of mechanisms is not an exhaustive one. Other mechanisms may (and probably do) exist and will fit in this division as well.

⁴ According to the conjunction of "folk" and "taxonomy", a folksonomy would seem to be in the gray zone already, but there is no restriction present in a folksonomy. It is simply a flat list of tags that users attach to all content. More quality is needed to move it to the gray zone.

⁵ <http://www.starlab.vub.ac.be>

⁶ <http://www.drupal.org>

⁷ http://en.wikipedia.org/wiki/Tag_cloud

⁸ Note that the topics RoadMap and RoadMap preparation are the largest in display. As we are currently in the process of construction a RoadMap at STARLab. The tag cloud correctly represents these as hot topics as a lot of people tag current posts in this area.



Fig. 2. From tag cloud to taxonomy in the Guide

relevant tags together (e.g., RoadMap and RoadMap preparation) and ordered them from generic to specific. Finally, we looked for even more generic terms to label the groups (e.g., Strategy). The end-result was a basic taxonomy that brings more structure to the Guide. The approach we use here is rather ad-hoc, and if different people (or even the same person at different times) create the taxonomy in this manner, we would end up with different results. However, for our current research the end-result is satisfactory. It is clear that the content will evolve continuously implying that the taxonomy will have to follow this evolution. This update will have to occur frequently, and as such, it is important that the construction of the taxonomy is kept as light (viz. neither complex nor time-consuming) as possible.

3.2 Guide Ontology

In order to take full advantage of all content present, we used the STARLab's DOGMA [24,25] approach to build a basic ontology for our Guide. This ontology captures the meaningful relations between all information objects in the Guide. Figure 3 displays the Guide ontology in NORM tree representation [26]. The formalization of the meaning in the ontology will allow us to perform reasoning and provide easy rule creation. For instance, because the system knows that Post causes Comment and that Post is categorized by Tag, we can easily add a rule stating that if Comment 'C' belongs to Blog Post 'BP' and Blog Post 'BP' is categorized by Tag 'T', that Comment 'C' is categorized by Tag 'T' is valid as well. If we combine this with knowledge present in the taxonomy, we can for instance find from Forum Topic 'id54' is categorized by Tag 'Dogma Studio', that Forum Topic 'id54' is categorized by Tag 'Development' is true as well.

This ontology describes the Guide overall content system. As this is relatively static⁹, the ontology will also be static. We foresee that we will have to update the ontology rarely.

⁹ Relative, as new types of posts (e.g., events) can be added in Drupal using modules.

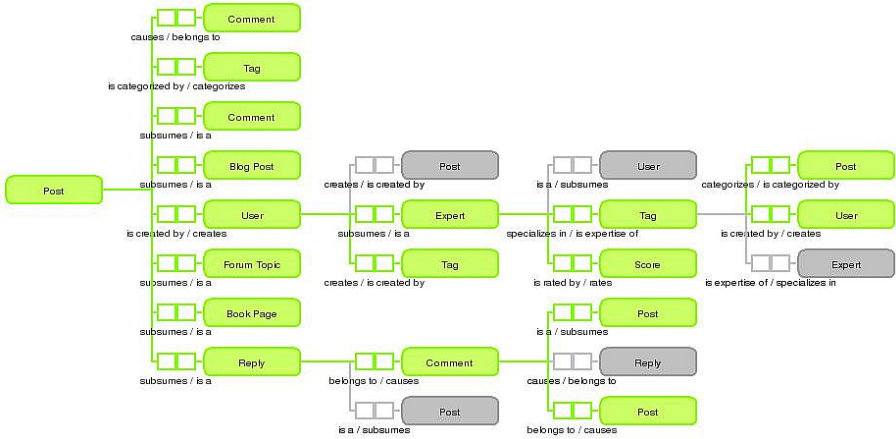


Fig. 3. Guide ontology in NORM tree representation. Gray-colored concepts represent duplicate occurrences appearing in the browsing process. These are included in the visualization in order to create a complete local context of the concept in focus.

3.3 Example Application

In the previous two subsections, we described how we enriched our Guide with both a taxonomy (to bring structure to the user-created tags) and an ontology (to describe the structure of the Guide itself). Now that this extra meaning is attached to the content, we must find ways in which we can usefully *apply* it. This way we can deliver valuable extra functionality to the users. We will demonstrate how we can achieve expert-group construction through our enriched Guide.

Suppose a new researcher starts working at STARLab. She is very interested in learning all about STARLab’s previous, current and future work. The Guide holds all this information, but despite¹⁰ both the tags and the structured posts (forum, book pages, ...), she feels a little unsure about where to start. She would like to get in touch with the right people, to ask the right questions and get the right answers. She navigates to the Expert Finder page. The new researcher is asked what she would like to know more about. She types ‘Development’, because she wants to talk to the people who built the tools for full understanding. The Guide now searches all posts that have been tagged ‘Development’. This search results in too few posts, so the system adds the tags ‘Dogma Studio’ and ‘Concept Definition Service’ as these are categorized under ‘Development’ in the current taxonomy (see subsection 3.1). The Guide now looks at the people who *created* these posts and at the people who *commented* on these posts (see the rule example in subsection 3.2), and ranks them as experts according to their activity in these posts. The system sets up a new forum topic saying that there is a new researcher looking for information about development in STARLab. It invites the

¹⁰ Or maybe because of the amount of structure and tags, as a large amount of such information is just as difficult to handle as a large amount of actual content.

two highest ranked experts¹¹ by email to discuss development issues with the new researcher. As soon as she determines that all answers are found, she can close this topic. If another new researcher comes by and looks for experts on 'Development', he will first be guided to the existing forum topic. If he does not find what he is looking for there, the process can start again.

The approach we followed and our example application can easily be transferred to other real-world problems. Consider the problem of integration of a community of immigrants into a native population. For these people, getting started is very difficult, as the amount of information is much greater than present in our Guide. The language aspect only adds to the complexity. A visit to the webpage of the town might help, but this community would benefit much more from a separate information source, targeted to their specific problems concerning integration, much like our Guide. The town's administration could offer a basic setup like a community portal, with some high-level structure (corresponding in general to the native information pages) and a less restricted part (like blogs). In cooperation with active individuals from the integrating community, this setup could be presented in the correct language. All community members are invited to join this portal and encouraged to locate and post solutions to issues there.

For instance, someone needs a form to indicate that he desires his administrative mails in French, rather than in Dutch. After several difficult visits to several administrative services, he locates the appropriate form and procedures. He posts the specifics of his administrative adventure on the portal (in a blog or specific forum topic) and tags it appropriately in his own language. On regular intervals, the folksonomy is converted into a taxonomy, which is coupled with the native town portal's information as well. Using functionality as described in our Guide example, the details on how to obtain and use the form can be located even more easily. The original discoverer of this information will be regarded as an expert. This approach will encourage people to use the system, and as such, make it usable and turn it into an active community *driver*.

4 Conclusions and Future Work

In this work we gave a brief overview of several meaning mechanisms and what their main advantages are. We divided these in free and restricted mechanisms and stated that free mechanisms tend to provide quantitative (but flexible) data, while restricted mechanisms could deliver more qualitative (but static) data. We identified the gray zone, which combines both sides and joins quality with quantity. Which side actually *seeds* the process of meaning generation is less important. We claim that there is need for processes and methodologies to support the continuous feedback loop legitimately. Both the community members and knowledge engineers must be active players in these processes in order to reach and benefit from the gray zone of meaning mechanisms. Only by combining quality with quantity in a legitimate manner can we achieve truly meaningful metadata. Meaningful not only for humans or machines, but for *both*. The example based on our Guide shows that it is indeed beneficial to move into the gray

¹¹ Note that these two people might not be regarded as the best development experts by their colleagues, but their number of posts indicate that they might be the most *helpful*. In the end, this is what is important for our current search.

zone. Using only basic meaning mechanism collaboration, we transformed the tags into a taxonomy and combined it with an ontology. The end-result was a system that provides a lot of possibilities for empowering communities. We gave a brief example on how to apply our approach in other, more significant real-world problems.

Future work should focus on this gray zone, and methodologies and mechanisms that thrive there. We described research that already entered this area, and we feel that these results provide motivation and grounds for more research. In our own work, we will focus on how we can improve our Guide example. Our current approach was rather ad-hoc (e.g., conversion of tags into taxonomy). We need to research further how we can turn this into solid meaning formalization and negotiation. Furthermore, we have to look at ways to create even more integration between the different mechanisms. We will also have to keep working on how we can then bring all this to actual application, and how this approach can benefit community members.

Acknowledgments. The research described in this paper was partially sponsored by the EU IP 027905 Prolix project and the Leonardo B/04/B/F/PP-144.339 CODRIVE project. We would like to thank our colleagues and Tanguy Coenen, Céline Van Damme and Eiblin Matthys from MOSI (www.vub.ac.be/MOSI) for their interesting feedback and discussions.

References

1. Robert Blumberg and Shaku Atre. The problem with unstructured data. *DM Review Magazine, February 2003*, 2003. http://www.dmreview.com/article_sub.cfm?articleId=6287.
2. Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American* 284(5), pages 34–43, 2001.
3. Aldo de Moor, Mary Keeler, and Gary Richmond. Towards a pragmatic web. In *Proc. of the 10th International Conference on Conceptual Structures, (ICCS 2002), Borovets, Bulgaria*, Lecture Notes in Computer Science. Springer-Verlag, 2002.
4. M.P. Singh. The pragmatic web. *Internet Computing Volume 6 Issue 3*, pages 4–5, May/June 2002.
5. Tim O'Reilly. What is web 2.0: Design patterns and business models for the next generation of software, 09-30-2005. <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>.
6. Gene Smith. Atomiq: Folksonomy: social classification, Aug 3, 2004. http://atomiq.org/archives/2004/08/folksonomy_social_classification.html.
7. Thomas Vander Wal. Explaining and showing broad and narrow folksonomies, February 21, 2005. http://www.personalinfocloud.com/2005/02/explaining_and_.html.
8. B.C. Vickery. *Faceted classification: a guide to construction and use of special schemes*. London: Aslib, 1960.
9. Barbara H. Kwasnick. The role of classification in knowledge representation and discovery. *Library Trends* 48(1), pages 22–47, 1999.
10. TR Gruber. A translation approach to portable ontology specification. *Knowledge Acquisition* 5(2), pages 199–220, 1993.
11. WN Borst. *Construction of Engineering Ontologies*. Centre for Telematica and Information Technology, University of Twente. Enschede, The Netherlands, 1997.

12. Asunción Gómez-Pérez, Oscar Corcho, and Mariano Fernández-López. *Ontological Engineering*. Springer-Verlag New York, LLC, 2003.
13. Eric Miller and Frank Manola. RDF primer. W3C recommendation, W3C, February 2004. <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>.
14. Frank van Harmelen and Deborah L. McGuinness. OWL web ontology language overview. W3C recommendation, W3C, February 2004. <http://www.w3.org/TR/2004/REC-owl-features-20040210/>.
15. John Ajita and Dorée Seligmann. Collaborative tagging and expertise in the enterprise, WWW2006, Edinburgh, UK.
16. Anthony C. Bloesch and Terry A. Halpin. Conquer: A conceptual query language. In *ER '96: Proceedings of the 15th International Conference on Conceptual Modeling*, pages 121–133, London, UK, 1996. Springer-Verlag.
17. Clay Shirky. Ontology is overrated: Categories, links, and tags, 2005. http://www.shirky.com/writings/ontology_overnated.html.
18. Ikujiro Nonaka and Noboru Konno. The concept of ba: Building foundation for knowledge creation. *California Management Review Vol 40, No.3*, Spring 1998.
19. Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions, WWW2006, Edinburgh, UK.
20. Patrick Schmitz. Inducing ontology from flickr tags, WWW2006, Edinburgh, UK.
21. Yannis Tzitzikas and Anastasia Analyti. Mining the meaningful term conjunctions from materialised faceted taxonomies: Algorithms and complexity. *Knowledge and Information Systems 9(4)*, pages 430–467, 2006.
22. Judit Bar-Ilan, Snunith Shoham, Asher Idan, Yitzchack Miller, and Aviv Shachak. Structured vs. unstructured tagging - a case study, WWW2006, Edinburgh, UK.
23. A. de Moor, P. De Leenheer, and R.A. Meersman. DOGMA-MESS: A meaning evolution support system for interorganizational ontology engineering. In *Proc. of the 14th International Conference on Conceptual Structures, (ICCS 2006), Aalborg, Denmark*, Lecture Notes in Computer Science. Springer-Verlag, 2006.
24. P. De Leenheer and R. Meersman. Towards a formal foundation of dogma ontology: part i. Technical Report STAR-2005-06, VUB STARLab, Brussel, 2005.
25. P. De Leenheer, A. de Moor, and R. Meersman. Context dependency management in ontology engineering. Technical Report STAR-2006-03-01, VUB STARLab, Brussel, March 2006.
26. Trog D. and Vereecken J. Context-driven visualization for ontology engineering. Master's thesis, Vrije Universiteit Brussel, 2006.

Virtual Individual Networks: A Case Study

Licia Calvi

Centre for Usability Research (K.U.Leuven) - IBBT
E. van Evenstraat 2A, 3000 Leuven, Belgium
licia.calvi@soc.kuleuven.be

Abstract. The paper will present the results of an empirical research dealing with the concept of virtual individual networks. This is a special case of social computing, in what it focuses on the concept of networking but from the point of view of an individual who is engaged in social relations (i.e., networking, indeed). The emphasis is also more on the social needs of individuals while networking and not so much on the technology *tout court*, although, of course, the identification of the individual social needs while networking will be used to define the requirements of the technology that might be used to this end.

Keywords: Social computing, communities, individual requirements.

1 Introduction

Everybody is, consciously or not, a member of several networks or communities: family networks, communities of friends, of colleagues, of people sharing the same interests and passions. These ties normally originate and are maintained with direct contacts, but the advent of new technologies makes the shift towards virtual (in a wide sense) contacts not only possible, but in some cases, when it enhances and strengthens them, even desirable.

The paper presents the results of an empirical research dealing with the concept of virtual individual networks. This is a special case of social computing, in what it focuses on the concept of networking but from the point of view of an individual who is engaged in social relations (so, in networking). The emphasis here is therefore mainly on the social needs of the individual who is engaged in networking with a view on the requirements of a possible technology that might be used to empower such networking. The identification of the individual social needs while networking in a real-life setting will be therefore used to define the requirements of this technology.

The paper is structured as follows. First, we will define the theoretical framework in which our empirical study was carried on, i.e., by giving a definition of communities and of virtual individual networks in particular. Then, the specific case study will be presented. Two hypotheses were formulated about social networking and about their consequences on the individual social life. They were tested by using a methodology that consists of a combination of *in vitro* and *in situ* techniques. The results of this experimental case will be discussed and they will be eventually generalized to support a wider community requirements modeling.

2 Virtual Communities¹

Different theories, several definitions, and various typologies of virtual communities exist (see, for instance [4], [5], [6], [7], [10]). However, we refer here only to one such formalization [10] because this is the one we have used to build the conceptual framework this research is based upon. And although there may seem to be no need for an additional model to explain virtual communities, we instead believe that a theoretical basis is necessary to frame and understand the study presented in this paper.

2.1 A Centripetal Movement in Virtual Community Development

According to Wang and al. [10], a virtual (or on-line) community can be defined by the intersection of three different factors. Such factors are place, symbol and virtual. They see them as the edges of a hypothetical triangle, meaning that they are equally important [10]. We follow this triple classification, but we rather see these elements as concentrically juxtaposed as they impact upon the notion of virtual community in a centripetal way (Fig. 1).

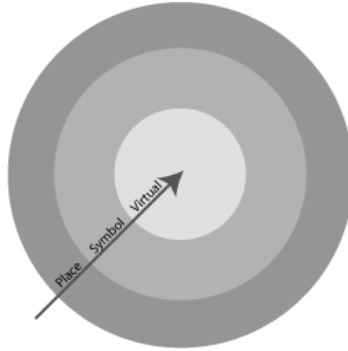


Fig. 1. Taxonomy of on-line communities

This movement starts at the most exogenous level, which corresponds to place. At first, indeed, a community is associated with a place because real communities are normally established in a physical location. Physical space is however not simply denoted at macro level, i.e., in the dichotomy urban/non-urban, but also at micro level, i.e., it can presuppose a further division within the urban concept of place towards any dedicated yet still physical place, like a sport club or a recreation centre, representing any activity which keeps several people together.

In this disaggregating movement, when the physical dimension is loosened, a community mainly rests on its symbolic connotation: its members ascribe a value and

¹ An earlier version of this part was presented at the CRESC Conference on Media Change and Social Theory, in Sept., 2006.

a meaning in being together, (see, also, in [8]), they identify the community as having a certain identity and create their own just by participating in this community life.

In its core, however, an on-line community is just virtual simply because it exists regardless of any location, i.e., in a non-physical dimension, as a bond among people that in many cases do not know each other personally (at least, at the beginning of the community existence), but who gather on-line simply because they have some common interest or they have some other form of relation they want to cultivate. Representations of the effect of this centripetal movement are for instance chatrooms, newsgroups, discussion lists, i.e., all forms of CMC (computer-mediated communication) environments that heavily rely on the spaceless interaction among their participants. These interactions may take different forms, and shape different kinds of communities: communities of people with the same interest who simply want to exchange information on it over one of another electronic medium; communities of people who share a phatic relationship; or communities of people who have a mainly work-related interaction. What remains in all these cases is then the relation among people. Such communities are then relational, to use a definition coined by Gusfield [2], and in antithesis to what he calls territorial communities in that they are location-unbound.

Although Wang and al. claim that “the virtual community exists in the mind of participants” [10: 411], some communities actually originates on-line and remain circumscribed to this space (we will call them *just-virtual*), while others are instead the extension of some communities already existing in the real world (we will call them *meta-virtual*) or yet others, starting from their on-line origin, even develop (with some or all of their members) a real, physical existence that progresses in parallel with the virtual, on-line one (we will call them *semi-virtual*). In this sense, communities clearly do not only exist virtually or in the participants’ head². An on-line community then emerges as its physical connotation has dissolved, when its symbolic and aggregating elements are still present although not predominant (and vary depending on the nature of the community itself) and if its distinguishing aspect is characterized by it being on-line.

2.2 Types of Relational Communities

Depending on how the relation has originated and further evolved, three types of communities can indeed be distinguished (see above). These three typologies can also be represented taxonomically (Fig. 2).

Figure 2 highlights the essential difference among these three forms of on-line communities:

- *just-virtual* communities originate on-line and mainly remain such. In this case, the fact that some of its members may decide to meet in the real world is seen more as the exception that confirms the rule: it happens but not so much.

² We have to specify here, however, that Wang and al. also use the term virtual as a synonym for real: “if one agrees that communication is the core of any community, then a virtual community is real whether it exists within the same physical locality or half a world away” [10: 411]. But our notion of reality implies the community being more linked to a physical existence.

Moreover, the outcome of transferring an on-line relationship in real life is not always successful. For the special case of chatters involved in a romantic, on-line relationship, for instance, statistics reveal that when the two people finally meet, in most cases the relationship is jeopardised and is then petered out (both in reality and on-line);

- *meta-virtual* communities originate in the real world but later further extend on-line;
- finally, *semi-virtual* communities originate on-line but, at a certain moment, also develop in the real world with a consistent number of their members (what is a major difference compared to virtual communities going off-line, as in the special example mentioned above), giving rise to a parallel real-world community. This is for instance the case of some expat communities. “Italians”³, for example, an on-line community of Italian expatriates, initially started as a virtual place where Italians abroad could meet and talk about their experiences as emigrants. It soon developed into a forum where also non-expat Italians and even non Italians (living both inside and outside of Italy) could meet. Additionally, a series of real events⁴ started to be organised where all community members could meet and get to know each other personally. These events still continue, regularly. Clearly, the side effect of these official gatherings is that people start to develop one-to-one or, also, more dedicated relationships with a restricted group of participants that they carry on both on-line and in the real world⁵.

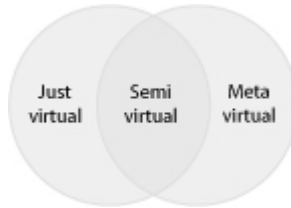


Fig. 2. Taxonomy of on-line communities

3 Experimental Study

We conducted a study on the importance of communities in everyday life and on the way people normally keep up with their friends and acquaintances that are within their social network. In this experimental study, the focus was on what we have identified above as meta-virtual communities, on their establishing and maintenance, that is on communities whose members share a relation which is based on some sort

³ <http://www.corriere.it/solferino/severgnini/>

⁴ The so-called “pizze Italians”, i.e., dinners to eat a pizza which are organised in different cities.

⁵ The difference between meta-virtual and semi-virtual communities is that for the former, the on-line existence is seen as an extension of its real world counterpart, and has therefore an ancillary role, whereas for the latter both existences, i.e., on-line and off-line, are equally important, although they might not involve all its members.

of common interest. The geographical character of these communities, although not really a *conditio sine qua non*, is certainly an important factor to keep the community alive and running.

More specifically, our study investigated the possibility for these geographical communities to become meta-virtual communities. This happens when people start to use technology to amplify the possibility of sharing information and of cultivating the same interest, something that until that moment has mainly occurred off-line.

As a side effect, we are also interested in studying whether communities which are based on the sharing of a common goal or interest by its members (e.g., organising a particular sport activity or sharing a passion for Austria, just to mention a few among the examples we came across) also imply a social bond among the members that goes beyond the declared objective of the community establishment and which may or may not be further strengthened by participating in a meta-virtual community life.

We formulated a hypothesis about the possible subjects who might be interested in being a member of a virtual (in-the-large) community. Our hypothesis presupposes that these people are already members of a community and foresees the possibility for this community to further develop on-line. In this sense, it deals with the notion of meta-virtual community described above (see Section 2).

This is a (persona) hypothesis (see Table 1). This states that everybody is, consciously or not, a member of several communities or networks. More precisely, the level of activity and the involvement within each of these communities vary depending on the individual's role in it and his/her motivation why being in it and, consequently, on the particular nature of the ties with part or all of the other members.

Table 1. Hypotheses under investigation

<i>Hypothesis 1</i>	Everybody is, consciously or not, a member of several communities or networks.
<i>Hypothesis 2</i>	The level of activity and the involvement within each of these communities vary depending on the individual's role in it and motivation why being in it and, consequently, on the particular nature of the ties with part or all of the other members.

In this analysis, we completely transcend from the discussion on-line vs off-line communities, i.e., over the existence of a continuum and over the blurring of a strict separation between these two entities, something that is going on in other domains (see, for instance, in [1]). Our assumption here is that in the case under investigation subjects are initially members of a community that exists in real life and we intend to investigate in how far this existing community may eventually be transported in a virtual/on-line setting and which with results. In order to do so, we first need to identify the subjects' requirements and necessities now, in this real community. Only in a second moment, we will try to transliterate such requirements into a virtual setting whose technical functionalities will have to be identified by then. To this end, a user and task analysis was carried on to identify the subjects' needs when involved in a community life (see next).

3.1 Experimental Protocol

The experimental protocol that was adopted for this user and task analysis is presented in details in the next subsections.

Our most important criterion in the selection of subjects was their being a member of a community, regardless of any other element. Since, as mentioned above, the focus lies on real-life communities (and on their possible extension and conversion into meta-virtual communities), we recruited subjects with an active involvement in a community life. By community, here, we refer to any *extra muros*, leisure association. So, among all the possible other communities everybody is involved in (like family, friends and work-related communities), we gave preference to those with a clear purpose, whose members are explicitly invited to contribute to the attainment of the community goal and which implement, to this end, a clear policy. Now, it is clear that any community is based on these three pillars, i.e., goal, members' motivation and involvement and internal policy (see also [8] for details). But we claim that in a network with family and friends all these elements are less prescribed and are mainly left implicit and tacit, while in an association with *extra muros* finalities they are normally very well formalized. Eventually, this is a formalization similar to that present in just-virtual communities, although the emergence of meta-virtual communities should entail the progressive adoption of those tacit and implicit rules that are also present in real-life, relational communities.

We tried to span over other factors, like education and familiarity with new media, in order to collect a heterogeneous study group. Like all these elements, age is also a transversal factor.

Additionally, we decided to introduce a further element of distinction among the subjects and divided them into two categories: individuals and couples. We recruited 10 individuals and 5 couples, but eventually received significant feedback only from 7 individuals and 2 couples.

This distinction was intended to highlight the possible *nuances* in intimacy relations. In particular, we were interested in investigating in how far the fact that at least one of the two partners is a member of an association and therefore participates in the association life may affect the intimate/family relationship and may, at the same time, shape the associative tie. This clearly refers to the second hypothesis mentioned above (see Table 1).

In our second group of subjects, however, there is a couple whose partners are both members of the same associations and there is a case where only one of the two partners is member of some associations. This group gives us the possibility to verify whether the bond among association members transcends the associative life to become something more intimate and if this intimacy is then extended to other people (in this case, the life partner) outside the association itself.

It should be clear that by focusing on members of associations, we intend to concentrate on people who have a well defined and definite community life which is more than just based on family ties or on loose friendship. The associations we came across range from walking clubs to any form of sport club, from an Austrian fan club to the association of the engineers of the University of Leuven, from a dance group to a nature guide association, just to give a few examples.

3.2 Methodology Definition

The methodology that was adopted to analyse the community life of the experimental subjects consisted of the combination of two different design methods:

1. a structured interview in the form of a questionnaire that was used to highlight the present needs and habits in terms of communication modalities and practices of the subjects. This questionnaire aimed at discovering, for each of the selected subject:
 - a. the type of most used communication technology;
 - b. their level of expertise and of confidence in it;
 - c. the reason for use, i.e., what, for which purpose, with whom, how frequently;
 - d. comments on the usability, cost/effectiveness, speed and simplicity of use of current technologies;
 - e. the importance of privacy issues;
 - f. a series of *desiderata* about current technologies and possible technologies to use;
2. a structured diary the subjects were asked to keep during a full week. In this diary, they had to report per day how information within the association was exchanged. In particular, they were asked to indicate with whom, for how long, what about, how and why they were communicating. The diary could be integrated with pictorial or printed material showing the life of the association: so, for instance, some of the respondents have been taking pictures of meetings or of events organised with the other association members; some others have glued in the diary emails exchanged within the associations; others have included newsletters or any other leaflet or flyer printed by the association to advertise its activities.

4 Discussion of the Results

Results have been processed in two ways:

1. quantitatively⁶, for the ones concerning the outcome of the questionnaire;
2. qualitatively, in a sort of text mining, for what concerns the diary.

If we look at the first ones, we see that our subjects are not particularly familiar with new media. They communicate with the PC, mainly using email, but mostly prefer the fixed and the mobile phone, which are both essentially used to make arrangements/appointments or to contact friends, although this kind of communication is not very frequent (a few times per month). The telephone is also preferred as a means to receive information.

If we try to correlate this mainly quantitative information to the qualitative one reported in the diary, we indeed notice that when communicating with the other members of the association the telephone is normally used. Sending SMS can be an

⁶ The actual quantitative results in graphical form are not included here due to a lack of space. They can be found on the project site: <https://projects.ibbt.be/vin/>.

alternative, but only in the case of brief messages, or as a confirmation/annulment of a previously taken agreement (which was normally taken by telephone). Direct contact via voice (through fixed or mobile phone) is also reported to be preferred for more intimate contacts (i.e., friends or family).

Emails are adopted when practical information needs to be exchanged (e.g., the programme of an activity), or when more people (e.g., all association members) need to be contacted at the same time.

From the diaries we can also infer that only in a few cases some of the members of the association are or become also friends. In some cases, they were friends to start with (and this may be the reason why our respondents were involved in that particular association originally), they never seem to become more intimate as a consequence of being members of the same association. In one specific case, this friendship extends to the life partners of the members and is responsible for extra associative activities (e.g., a dinner for a member's birthday). But in this particular case, our subject is a member of the association of the engineers of the University of Leuven, i.e., a group of people who graduated in the same year and who are therefore long-course friends. This result is therefore not significant for our analysis.

Of the two couples that responded, one represents the case where both partners are members of the same associations. Moreover, one of these associations is chaired by the husband's brother and is a neighbourhood committee that organises on voluntary basis activities for the neighbourhood (this may refer to Oldenburg's notion of 'third place', see in [9] for details and also in Section 6). In this case, then, the relationships with the other members are also very intimate and involve both partners. Of the other two associations these subjects are still members, one is a very huge sport club (more than 800 members); the other is the association of the Austria lovers (with more than 100 members) which organises a monthly activity in the form of a walk and a year activity in the form of a trip to Austria. In both these cases, then, the relationship with the other members is mainly formal and not particularly intimate.

5 General Considerations

By abstracting from the results mentioned in Section 4, four more general principles may be drawn:

1. *The computer is a tool and not a medium, and it is certainly not a goal in itself.*
The computer (as a synonym here for a CMC context) is not yet integrated into the daily activities of the respondents but it is merely used for a specific function (i.e., to confirm an appointment, to distribute some material to more people, and this all via email). Moreover, they do not seem to appreciate and even to understand the potential of CMC environments, the pleasure of being on-line and meet new people, share with them their interests and discover new information.
2. *Most people are not pioneers in the discovery of new communication media.*
The average, common person seems to be more inclined to take over trends and habits when they are established rather than to anticipate them. From the results mentioned above, we experienced a general need to see things (in this case, advancements) happen rather than a willingness to make them happen. It is

difficult to have a long-term vision, to see beyond contingent practices, maybe because of the too fast changes in communication media.

3. *Whatever new element, habit, or activity is taken over, this has to be adapted into the frame of a well-established habit.*

Any modification in one's life, even any improvement, has to fit to one's established set of habits and practices and not be perceived as an abrupt intrusion in one's routine. Such modifications must take place gradually and must be supported by a more general shift in practices occurring in the social environment of the individual. Often, in order to be effective, training is necessary to convince the individual of the need and the benefit of a certain change in habits (see also in [3] and [9]).

4. *There may still be some form of reluctance and even fear in accepting unknown practices.*

This principle seems to summarise the previous ones. Any new practice has to be brought closer to the people who will have to adopt it and explained to them in their own language. They have to feel that this will represent an improvement and not an obstacle for the achievement of their goals and the fulfillment of their lives.

6 Conclusion

The purpose of this study was to identify the way in which community communication is carried on in real-life settings in order to infer how this could eventually be transferred and translated into an on-line virtual environment. We have been referring to this evolution as a case of meta-virtual community. The results presented in this paper seem to suggest that within an association members mainly maintain a formal relationship, which is essentially devoted to the exchange of information relative to the association activity or to the pursuing of the association goal.

Several other issues are related to this. They include, for instance, issues of intimacy and of psychological closeness. We have tried to tackle them by analyzing how ties evolve and if they evolve (from less close to more intimate), and in how far relationships within an association may affect the intimate relations inside a couple. To this respect, we have identified different possible settings, all of them with diverse implications for the development of the community.

For all these social needs, however, not many technological means are used and, by extension, not many functionalities may be needed when integrated into or fostered by some form of social software tool. We have indeed derived some more general principles by abstraction from the experimental results which show an overall tendency towards the maintenance of established and conventional habits which prevents people from exploring and being keen on adopting new practices.

The key to a change in this direction might consist in persuading people that CMC can become an integral part of their life, and that this integration can positively affect all their life. A concrete way of doing so may take the form of those 'virtual third places' described by [9] where the shift towards a virtual community *tout court*, i.e., according to the definition given in Section 2.1, is gradual because the link with the

real-life community is more evident. This is precisely what the meta-virtual communities we have been investigating here are meant for.

Acknowledgments. This research is part of the IBBT ISBO-VIN project. Members in the consortium are: CUO (K.U.Leuven), SMIT (Free University Brussels), MICT (Ghent University), ICRI (K.U.Leuven), COSIC (K.U.Leuven), EDM (Hasselt University), ETRO (Free University Brussels), IBCN (Ghent University), DESICS (IMEC).

The author would like to express a special thank to the partners from the social science institutions for the numerous discussions and reflections that have helped her frame her research.

References

1. Frangoulis, A., Calvi, L.: *Virtuality and Multiplicity of Contexts: Social Implications*. CRESC Conference on Media Change and Social Theory, Oxford (2006)
2. Gusfield, J.: *The community: A critical response*. Harper Colophon, New York (1975)
3. Keeble, L.: Why create? A critical review of a community informatics project. *Journal of Computer-Mediated Communication*, Vol. 8(3) (2003)
4. Kim, A.J.: *Community building on the web*. Peachpit Press, Berkeley (2000)
5. Lazar, J., Preece, J.: *Classification Schema for Online Communities*. (1998). Available at: http://www.ifsm.umbc.edu/~preece/Papers/1998_AMCIS_Paper.pdf
6. Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., Cosley, D., Frankowski, D., Terveen, L., Rashid, A. M., Resnick, P., Kraut, R.: Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, Vol. 10(4) (2005)
7. Preece, J.: *On line communities: designing usability, supporting sociability*. John Wiley & Sons, Chichester (2000)
8. Preece, J., Maloney-Krichmar, D.: *Online Communities: focusing on sociability and usability*. In: Jacko, J.A., Sears, A. (eds.): *Handbook of Human-Computer Interaction*. Lawrence Erlbaum Associates Inc., London (2003)
9. Soukup, C.: Computer-mediated communication as a virtual third place: building Oldenburg's great good places on the World Wide Web. *New Media & Society* Vol. 8(3) (2006) 421-440
10. Wang, Y., Yu, Q., Fesenmaier, D. R.: Defining the Virtual Tourist Community: Implications for Tourism Marketing. *Tourism Management*, Vol. 23 (4) (2002) 407-417

Agent Community Support for Crisis-Response Organizations

Hans Weigand

Infolab, Tilburg University, P.O. Box 90153, 5000 LE Tilburg,
The Netherlands
H.Weigand@uvt.nl

Abstract. Crisis response organizations can be supported effectively by means of agent communities where agents represent human actors or organizational roles. An agent community can be organized in several ways. The paper defines requirements on agent community architecture and coordination structure from the point of view of crisis response, and proposes an architectural solution. Particular attention is given to the distribution of information.

Keywords: Crisis-Response Systems, Multi-Agent Systems, Communities.

1 Introduction

Crisis response organizations are shifting from hierarchical and static structures to dynamic networks. Modern information and communication technology enables the dynamic creation of ad-hoc networked organizations. This trend is similar to the evolution towards network-centric organizations in the military [1].

During the last decade significant progress was made in the development of integrated crisis response systems such as monitoring systems. Experimental computer supported cooperative work systems were developed e.g. for planning routes and coordinating crisis response teams. The introduction of GPS and GIS added significant functionality to the disaster monitoring and emergency team dispatching systems.

Multi-agent systems [11] have been suggested by several authors as an effective solution to solve the disaster situation management tasks due to the distributed organizational framework, the use of mobility of certain kinds of agents, and the fact that MAS supports smoothly the idea of a community of collaborating human and system agents.

The objective of this paper is to investigate the requirements on the support of crisis response organizations by means of multi-agent systems. Particular attention is given to the support of communities and to the distribution of information. Section 2 reviews some examples of MAS support for crisis response management. In section 3, we list general requirements, including community support, and underline the need for community support by a small case study. In section 4, we focus on the information and communication support and discuss a couple of options.

2 Multi-Agent Systems for Crisis Response Management

Multi-Agent Systems (MAS) have been suggested as a suitable solution for Crisis Response Management (CRM) systems. We summarize some of the work. Van Veelen, Storms and Van Aart [10] investigated several MAS coordination strategies from the point of view of agile crisis response. They distinguish between knowledge based coordination, such as the military SMDS systems, rule-based coordination typically based on negotiation in a market-like structure, and skill-based coordination in which there is no interaction between the agents, agents decide on their actions based on local optimization rules. They also refer to ant-based coordination as an alternative approach.

Jakobson et al. [5] extend a basic MAS with the capability of situation awareness. Central to this architecture is a Situation Model, a real-time constantly refreshed model of the disaster, on the basis of which relief operations can be planned. The idea is that agents are not only reacting to messages or single event notifications, but use event correlation: a conceptual interpretation procedure that assigns new meanings to a set of events that happen within a predefined time interval. The output can itself be used for further interpretation. This event correlation is realized by means of case-based reasoning techniques, where a case is a template for some generic situation.

A different application of MAS techniques can be found in the area of simulation. For example, [8] uses multi agent systems to simulate evacuations and to improve upon traditional crowd simulators.

3 Crisis Response Organizations – General Requirements

An example of a crisis response organization is a medical relief operation after a disaster that includes field mobile ambulatory aid, evacuation processes, emergency hospital operations coordination and logistics support for medical supplies. There may be several relief organizations participating, which may involve language and equipment differences. The scope of the disaster may put local medicine facilities out of order, and it may place relief teams in hardship conditions or at risk because of for instance limited food and water supplies and lack of law enforcement. Specific tasks that need to be supported by the CRM system include overall planning of the medical recovery effort (personnel, equipment, supplies), dispatching, scheduling and routing of mobile ambulatory and other emergency vehicles, evacuation of victims, maintenance and care of relief personnel, and communication and coordination between medical teams as well as to other relief operations [5]). What requirements does such a situation present to the crisis response organization and its CRM systems?

Agility and discipline. According to John Harrald, member of the US National Research Councils Committee on Using Information Technology to Enhance Disaster Management [4] crisis response should be both *agile* and *disciplined*. The ability to improvise, create, and adapt in the face of unforeseen events and circumstances has been the key to successful response and recovery efforts (e.g. the response to the 9-11 attacks, response to the 2004 Florida hurricanes). However, agility is not all. The international response to the December 26, 2004 tsunami shows that government and

non-government organizations can be extremely agile and creative in responding to a disaster of historic proportions. The lack of discipline, however, was evident in the lack of coordination and communication, the ad hoc mobilization of resources, ineffective use of technology, and inability to integrate diverse organizations.

Robustness. Another often-mentioned requirement on crisis response systems is *robustness*: it is typical for crisis situations that parts of the network may be malfunctioning and this should not disable the system as such.

Embedding. To deal with crisis response systems effectively when it is needed, actors should be familiar with the systems. This can be achieved in two ways. One is by means of education and regular training events. The other is by embedding the crisis response system in a system that the users also use for normal activities. How this is to be done depends on the particular situation. For example, security guards may use a mobile communication system for their daily work; the same systems can get more functions in the case of an emergency. Although the guards will have to know these extra functions by training, the fact that they are accessible with the same devices and a familiar user interface, will improve their ease of use.

Community support. In an emergency situation, such as the one sketched above, it typically happens that many groups have to work together that are not used to collaborate. To some extent, the collaboration can be improved by the use of common systems (standards) and combined training events. This should certainly be done, but there will always remain a high level of indeterminacy. We derive from this that CRM systems should, on the one hand, be effective in supporting communities or groups (such as a group of firemen, or a medical team), and on the other hand support agility in setting up connections between groups.

A community is “a group of people bound together by certain mutual concerns, interests, activities and institutions” [9]. When people are professionals (as in the case of crisis response) and the collaboration is mostly or completely enabled by information technology, it can be called a virtual professional community [7]. A crisis response organization will almost never be completely virtual, but the IT support is becoming more and more important.

Communities need to be supported. We mention a number of generic support instruments that are definitely relevant for crisis response organizations. First, a community needs a *door keeper* or guard that adds new actors to the community and can remove them. In this way, it can be traced down who is a member of the community and who is not. In the case of an agent-supported community, the door keeper can be a special agent that allows other agents (provides them with a proxy) on the basis of certain rules. This leads to the second generic instrument: *rules* (or norms, or institutions). Rules for admission, rules for communicating, rules for decision making, etc. These rules may evolve over time, hence they should preferably be made explicit. Thirdly, communities use *roles*; a role implies certain authorizations and goals [3]. A role can be fulfilled by various agents, or by different agents in the course of time. One advantage of the use of roles is that one can send a directive to a certain role without having to know which agent is fulfilling this role.

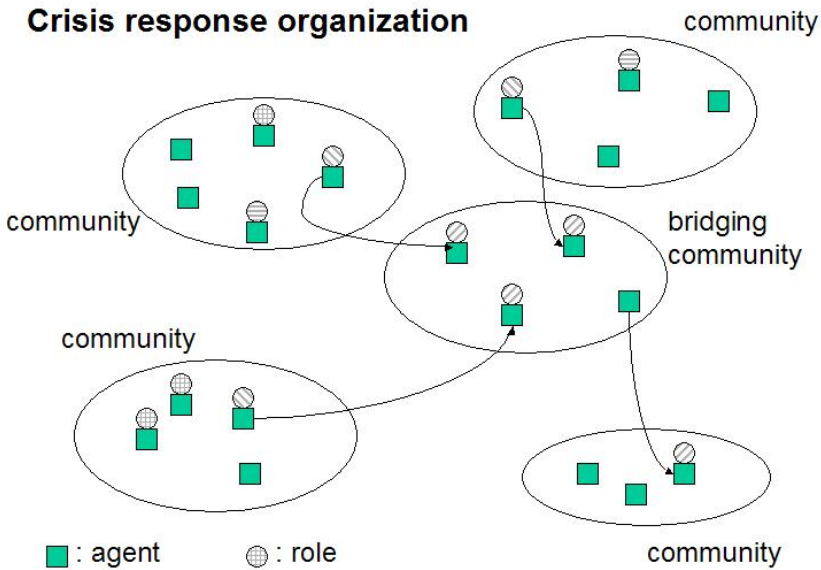


Fig. 1. Crisis response organization as a connected set of communities

As we said, a crisis response organization is typically a network of various professional teams. Each of these teams has its own discipline and way of collaboration – it would not be wise to destroy that structure and put all members of all teams together into one overall community. So we need not only support for intra-community collaboration but also for inter-community collaboration. However, we do not need completely different instruments for that. One way of supporting inter-community collaboration is by using *bridging communities*. A bridging community is a community (with door keeper, rules and roles) whose specific objective is to coordinate behavior between other communities. To this end, it allows one or more agents of each of these communities in the role of representative. These agents act as linking pins. Bridging communities can themselves be represented in other (higher) bridging communities (we put the word “higher” in brackets because bridging communities can be organized as a hierarchy, with one highest bridging community at the top, but this is not necessary). A crisis response organization can be defined now as a set of communities and bridging communities in such a way that there is a path (at least one) between any two communities in the set (roughly said, the set is a connected graph – see Fig.1).

A critical reader may object that we blur the distinction between the crisis response organization – the communities of human actors - and the crisis response management system – the multi-agent systems that assumingly support these actors. Of course, these are still different levels. However, in the case of MAS systems, the structures of the CRM system can reflect very well the organizational structure at the human level. And it is to be expected that in the future the support system will *be* the organization of the community, as it is already the case nowadays in many business organizations and networks.

Case study. The need for community support is underlined by a small case study that we have performed in the area of Incident Response. All the universities in the Netherlands have an incident response team to respond to network attacks, such as a worm virus, phishing, or hacking a server. At Tilburg University, the IR team consists of 7 members of which every week at least one is active. Usually, the active (responsible) member consults the other members when an incident occurs, but he is authorized to make unilateral immediate decisions if needed. Within the university, the IR team has to deal with the system administrators, as these are typically the people that have to take an action like disconnecting a machine from the network. It also has to deal with the police, in the (rare) cases that an incident happens that by law must be reported. The university IR team is linked to a national organization of universities (SURFNET) whose IR team called CERT-NL has a coordinating role and works on a 24-7 basis. This team can either forward information to other teams (e.g., there is a machine in your network that is sending spam mail) or respond itself. It is important to assess the possible impact of an incident: e.g., if a certain server software occurs to have a certain security hole in one place, then all the other installations of that software are vulnerable as well and need to be upgraded. Both CERT-NL and the university team participate in a national forum called GOVCERT.NL. Within GOVCERT, IR teams meet regularly for sharing knowledge and experience. On an international level, there are similar forums, most notably the worldwide forum FIRST in which GOVCERT participates. The forums have an important social networking function: the people that meet there regularly get acquainted with each other, and during a crisis situation it becomes much easier to ask help. Furthermore, the forums provide knowledge to their members in the form of best practices and historical archives of advisories.

The case study illustrates a crisis response organization whose crisis management is embedded in a pre-existing organization and organizational network. The community aspect of the teams and forums is evident, and clearly essential. Some processes that are currently performed by traditional instruments such as email could profit from more intelligent tools such as agent support and knowledge management. However, such tools should never replace face-to-face meetings.

4 Crisis Response Management – Communication and Information

At the information level, crisis situations put severe demands on the distribution of data across teams of people and systems, and the ongoing data collection and changing state makes the overall picture very dynamic. There is a strong benefit to the overall effort when different teams can share relevant information. However, this does not mean that the more information is disseminated the better. Firstly, people can be overloaded with information in such a way that they may miss the really relevant items or do not have time anymore to do their jobs. Secondly, not all information may be interpreted correctly by all actors involved in their different communities, which may cause confusion. Thirdly, there is the danger of spreading rumours and other kinds of uncertified data, which again may cause confusion and even panicking. So what are useful mechanisms to support the communication and information

dissemination within and across communities in a crisis response organization? We assume that the CRM takes the form of a network of MAS's.

Pre-defined workflows and event notifications. A crisis situation requires agility and ad-hoc solutions. In that respect, pre-defined workflows have limited value. However, discipline is also important. A professional crisis response organization and its communities will have certain structures in place, and these can be supported well by having important workflows and event notification schemes defined and deployed. Evidently, these structures should not exclude other ways of communication, of circumventing the system.

Global situation models. It is not wise to distribute all information immediately to all actors involved, but what is important is that all actors can access all information, if needed, and develop a common understanding. The task of maintaining a global situation model can be assigned to one actor (or team of actors) with a special agent. The model should be accessible in multiple ways (directly by other agents, or via a web page to human actors with a computer or PDA, or via the human actor). To support access by other agents (from different communities), a bridging community can be defined that contains the global situation agent plus representatives of the various agent communities. Such representatives we call *information agents*. It is not necessary that all communities have their information agent directly connected to the global situation agent; there may be one or more information agents in between. What is important is that all communities are connected directly or indirectly to the global situation agent. Preferably, the connection has a certain redundancy (robustness).

The task of maintaining a global situation model is not trivial. It requires interpretation, sometimes pruning, prioritization, and in general improvement of the quality of the data. The detailed design of a global situation model agent is not in the scope of this paper.

The installation of a global situation model team may take some time. In the mean time, what to do? In a professional crisis response organization, we should expect each community (each participating team or organization) to have its own situation manager (and situation manager agent). When different communities connect into a crisis response organization, these local situation managers can hook up in a peer-to-peer fashion (using dynamically set up bridging communities), and exchange their data. Once a global situation manager is in place, these local situation managers turn into information agents that report to and acquire information from the global situation agent.

In the above, we hinted at the problem of uncertified data spreading around. Hence we propose to make use of a confirmation system. Each data item should not only have a propositional content (the core information item, e.g. "the road between X and Y is blocked") but also a modality including the source, a time stamp, and the number of independent confirmations (perhaps distinguishing between authorized and unauthorized confirmations). The system should encourage actors and agents to confirm data that they receive if indeed they have independent evidence. A human actor should be able to make such a confirmation in the most convenient way (one click), after which the confirmation is forwarded automatically via the information agents to the global situation manager.

Rule-based interpretation. A crisis response organization can and should have a global situation manager, but this does not contain of course all the knowledge from the various teams. Knowledge is distributed. We cannot expect that all information is communicated as clear action instructions simply to be performed. Agents should be equipped with *interpretation rules* that may fire on the occurrence of certain events (usually incoming messages) or the occurrence of complex data patterns [5]. These interpretation rules can be complemented with action rules that on the basis of interpretation results undertake certain actions, such as notification of the human actor. Preferably, these rules are easy to add and also to exchange between agents, so they should be treated as first-class objects.

Problem patterns, solution patterns. If rules can be exchanged, then it is also worthwhile to collect and consolidate rules that have proven to be useful. This is typically not done during the crisis situation, but afterwards, and before a new event. It is to be expected that certain agencies, such as national defense organizations, will build up knowledge bases of problem patterns (for the recognition of a problem) and solution patterns (for a heuristic solution to a certain problem) that can be imported by crisis response organizations when needed. A disaster plan is an obvious example of a solution pattern, but there may be many more. For example, a medical team may suddenly be confronted with hostage taking of one of its members by some violent group, and may not know how to respond to that. Finding the right patterns for such a situation may itself be a hard information task. To improve recall and precision, it might be useful that the search request is supported by a context description such as maintained by the local situation manager/information agent.

Overhearing. In recent agent research, it has been argued that group cooperation can benefit greatly from so-called overhearing [2]. For example, a person asks help from another team member and does it in such a way that the whole group can hear it. Then it may happen that a third person in the group overhears the conversation and provides unsolicited help, as he happens to know the answer or something relevant for solving the problem. To make overhearing effective, it must be assumed that the group has a shared model (so that the overhearer can understand what he observes), that the communicating agents are willing to receive unsolicited help and hence make their behavior public. This typically means that the agents in the community are supposed to be cooperative and benevolent.

The principle of overhearing is applied, for example, in the IRT case described earlier. When an active member of a IR team sends an email to someone (e.g. a system administrator), this email is automatically cc-ed to all members of the team. They don't need to react, but if they think it is appropriate, they can come up with suggestions.

Imitation. In human society, imitation is a powerful instrument of social coordination, but it has not been explored very much yet in computerized systems and MAS – except perhaps in the form of particle swarm algorithms [6]. Imitation could be explored in several ways. In the situation that there is one experienced actor who knows what to do and less experienced actors that don't, it should be possible to set up a “follow-me” relationship. This assumes that the follower can observe the

behavior of the model. A possible example is geographical routing where actors (or their agents) automatically give off data about their position and direction, e.g. by an in-built GPS. If the data is automatically given off, the model may make his behavior observable by recording his actions into his agent (speech would probably the most convenient form of recording, possibly combined with automatic recognition and digitalization). The follower puts his agent into follower mode so that he gets the recorded information. This may be real-time, but it should also be possible to retrieve it later (in other words, the recordings should be stored).

A weak form of imitation is flock behavior, which can be very beneficial. To support that, all actors in the community should record their doings (or some of them), and broadcast these to the other members' agents. This information is not forwarded to the human actor, but analyzed by the agent himself. Analyze for what? One interesting question is whether the actor is deviating from the other ones (moving away from the flock – literally, in the geographical sense, or in terms of the kind of behavior). That could lead to a warning signal to the actor. Another objective of analysis is comparison: are the other ones doing better? In reaching a certain place, or in successful action (e.g. number of patients helped). If so, again the actor may be signaled, or be put into follow-me mode with a more successful member.

5 Conclusion

In this paper, we have looked into a specific kind of community, that is, communities involved in a crisis response organization. It has been suggested that such a community can be supported quite well with a MAS-based Crisis Response Management system. It has been argued that a Crisis Response Organization is to be regarded as a connected set of communities, and that therefore community support is one of the requirements on a CRM. Some minimal ways of community support have been discussed. Special attention has been given to the dissemination of information within the Crisis Response Organization and its communities. A couple of instruments have been discussed that can be used separately but preferably in combination, ranging from traditional workflow management solutions to more advanced mechanisms, such as imitation, that still need to be explored. Although the instruments were presented here specifically for crisis response management, they may be useful for other kinds of communities as well.

References

1. Alberts, D.S., Gartska, J.J., Stein, F.P.: Network Centric Warfare: developing and leveraging information superiority. DoD CCRP (2002) Available at <http://www.dodccrp.org>
2. Busetta, P., L. Serani, D. Singh, and F. Zini: Extending multi-agent cooperation by overhearing. Proc. the Sixth International Conference on Cooperative Information Systems (CoopIS 2001), Trento, Italy (2001)
3. Dastani, M., V. Dignum, and F. Dignum: Role-assignment in open agent societies. Proc. AAMAS'03, Melbourne, ACM Press (2003), 489-- 496

4. Harrald, J.R.: Supporting agility and discipline when preparing for and responding to extreme events. ISCRAM 2005, keynote speech (2005)
5. Jakobson, G., N. Parameswaran, J. Burford, L. Lewis, P. Ray: Situation-aware Multi-Agent System for Disaster Relief Operations Management. Proc. ISCRAM 2006 (B. van der Walle, M. Turoff, eds). New Orleans (2006)
6. Kennedy, J. and R.C. Eberhart: Swarm Intelligence, Morgan Kaufmann Publishers (2006)
7. Moor, A. de: Empowering Communities / a method for the legitimate use/driven specification of network information systems. Dissertation, Tilburg University (1999)
8. Murakami, Y. Minami, K. Kawasoe, T. Ishida, T.: Multi-agent simulation for crisis management. IEEE Workshop on Knowledge Media Networking (KMN'02), Kyoto (2002)
9. Talbott, S.: The Future Does Not Compute: Transcending the Machines in Our Midst. O'Reilly & Associates, Inc. (1995)
10. Veelen, J.B. van, P. Storms, C.J. van Aart: Effective and Efficient Coordination Strategies for Agile Crisis Response Organizations. Proc. ISCRAM 2006 (B. van der Walle, M. Turoff, eds), New Orleans. (2006)
11. Wooldridge, M.: An Introduction to Multi Agent Systems. Wiley & Sons, Chichester, UK (2002)

Aggregating Information and Enforcing Awareness Across Communities with the Dynamo RSS Feeds Creation Engine: Preliminary Report

F. De Cindio¹, G. Fiumara², M. Marchi³, A. Provetti²,
L.A. Ripamonti¹, and L. Sonnante⁴

¹ DICO, Università degli Studi di Milano

² Dip. Di Fisica, Università degli Studi di Messina

³ DSI, Università degli Studi di Milano

⁴ Fondazione Rete Civica di Milano

Abstract. In this work we present a prototype system aimed at extracting contents from online communities discussions and publishing them through aggregated RSS feeds.

The major foreseeable impact of this project to the Community Informatics field will be helping people to manage the complexity intrinsic in dealing with the huge amount of dynamic information produced by communities, in particular, keeping up with the evolution of several simultaneous discussions/information sources.

A special version of the Dynamo system, which is described here, was deployed to endow RSS channels to the forum of the Milan Community Network (RCM).

Keywords: community informatics, on-line community, community network, knowledge management, artificial intelligence, knowledge sharing, RSS, XML.

1 Introduction: Human Attention in the Modern Time

Already in 1971 Herbert Simon was envisioning the advent of the so-called “attention economy”, claiming that “...in an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it” (Simon 1971, pp. 40-41).

Nowadays his assumptions have proven to be farseeing, since the “attention economy” has become the everyday reality we are living in. Undeniably, one among the most time-consuming human activities in modern times is keeping up to date with a huge amount of continuously evolving and changing information, conveyed to us through a mix of different multimedia source, whose evolution too, is a restless process. As a consequence of this overexposure, people’s attention has become one among the rarest and most valued resources.

Business world (and especially researchers and practitioners involved in the organizational, human resources management or marketing disciplines) are well aware of this phenomenon, and are struggling to define how to appropriately deal with it: let's think of the growing interest around virtual and networked organizations (Lipnack & Stamps, 2000), communities of practice (Wenger, McDermott & Snyder 2002), or tribal marketing (Kozinets, 1999 or Cova, 2003), but it pervades every aspect of our lives, in every moment, being it a working or a leisure one.

The complexity intrinsic in dealing with such information overload grows exponentially for people actively involved in on-line communities, since they must (or want to) keep the pace of discussions that may be very "active" and even split across multiple threads, or – in the worst case – among different (sub)communities. This means not simply adding one (or more) information source, but also being able to manage the emotional involvement and the selection of the highly-prized relevant pieces of information in the sea of chatters that form the humus upon which the collective intelligence that creates knowledge in the social network buds (Polanyi, 1967 Wenger, McDermott & Snyder 2002, Armstrong and Hagel, 1998).

This escalation in the amount of information we [are forced to] process every day underpins the skyrocketing evolution of ICT (Information and Communication Technology), that continuously makes new interaction media available, and that has been paralleled by a similar evolutionary process in the ways people are using (and from time to time also twisting) them to support social interactions through innovative paradigms, enabling more effective informative and emotional exchanges. Perhaps the most outstanding demonstration of this process is the increasingly successful "podcasters¹ movement."

2 RSS, ATOM, Aggregators and Other "Exotic Technologies"

One relevant step in the direction of helping people managing such complexity has been made by Really Simple Syndication (RSS) and ATOM readers, that enable users to collect simultaneously news from different selected sources. RSS readers completely revolutionized the paradigm according to which people collects information from the news published on the Internet: it is no longer the user that searches the information she is looking for, but it is the information she values that reaches directly its "consumers", by downloading on their devices (PC, iPod, etc.).

¹ **Podcasting** is the method of distributing multimedia files, such as audio programs or music videos, over the Internet using either the RSS or Atom syndication formats, for playback on mobile devices and personal computers. The term **podcast** like 'radio', can mean both the content and the method of delivery. The host or author of a podcast is often called a **podcaster**. Podcasters' web sites may also offer direct download or streaming of their files; a podcast however is distinguished by its ability to be downloaded automatically using software capable of reading RSS or Atom feeds.

Usually a podcast features one type of 'show', with new episodes released either sporadically or at planned intervals such as daily or weekly. In addition, there are podcast networks that feature multiple shows on the same feed (from Wikipedia – www.wikipedia.org).

Unfortunately this solution is still too basic, since RSS readers offer a very limited interactivity to users, that are enabled uniquely to subscribe the specific “feeds” they are interested into. Clearly they are very helpful, avoiding people to spent an amount of time daily to search for the information they need across scattered sources, but they are unable to select only “really relevant” information or to aggregate them consistently with their semantic meaning.

If this limit is quite tolerable while reading feeds collected from “news” (e.g. an on-line newspaper) – after all I can quite easily skip irrelevant news by reading their title -, the problem assumes different boundaries when applying RSS readers to on-line communities. In this latter case no editing service exists to format information according to a uniform standard; on the contrary, information flows freely in the stream of the discussions, and people posts their opinions and messages according to their mood, not worrying about how their posts will look like if collected by a RSS *aggregator*². According to this paradigm, it is quite impossible for the user to distinguish between two (or more) different replies made by different community members to the same post, unless reading the whole contents. This may add a dramatic overhead, that risks to reduce to zero the advantages offered by the RSS reader (after all, what is the utility of downloading on my PC exactly the whole discussion taking place on my favoured on-line community instead of reading it on-line?).

To address this problem we are now working on the Dynamo project, which is aimed at developing an RSS feeds creation engine based on artificial intelligence techniques – namely the Answer Set Programming (ASP) - in order to provide community members with sophisticated news digests, tailored on their specific information needs.

At the moment, a special version of the Dynamo system is undergoing a preliminary phase of testing in several forum of the Milan Community Network (RCM: *Rete Civica di Milano*, cfr. De Cindio et al., 2003).

This article is organized as follows: after a brief survey of news syndication techniques we will give an account of the RCM community network, where the Dynamo content extraction technique, which is described next, has been applied.

3 Starting Point: The Current RCM Technological Platform

RCM is a network of more than 20.000 citizens living in the Milan area, forming a community network (CN – cfr. De Cindio & Ripamonti, 2006), whose main focus is “being a citizen of Milan” (in the broadest sense of the term). RCM has been founded in September 1994, as an initiative of the Civic Informatics Laboratory at the University of Milano, and, since then, it has developed several projects aimed at designing and implementing services for citizens, local no-profit associations and local small businesses.

² An **aggregator** or **news aggregator** is a type of software that retrieves syndicated Web content that is supplied in the form of a web feed (RSS, Atom and other XML formats), and that are published by weblogs, podcasts, vlogs, and mainstream mass media websites.

From a technological point of view, RCM core community services are currently managed through the *FirstClass* server, produced by OpenText Corp.

FirstClass is a messaging and communications platform, mainly addressed to schools, learning organizations and businesses, but also to on-line community and CNs. FirstClass provides its users with the ability to communicate and share resources and information via email, conferencing, directories, individual and shared calendars and on-line chats. All these features can be accessed both through a Web interface and a proprietary client.

In order to understand RCM technological choice, it's important to bear in mind that, when RCM started in 1994, dialog and communication facilities through the Internet were not user-friendly enough for “everyman” (at that time they were implemented on text-based BBS (Bulletin Board System), and the World-wide Web was in its early stages, still not providing interactivity). In that technological landscape, FirstClass was, and to some extent still is, among the solutions with the best trade-off between costs and requirements. Indeed, FirstClass:

- is highly interactive, hence good at supporting active citizens producing contents of local interest;
- is endowed with an easy-to-use windows-based interface;
- supports a wide range of communication protocols;
- supplies a light “client application”, that guarantees reasonable performances even with cheap/out of date personal computers and modems (thus being an affordable mean of communication for a large group of potential users).

At present, RCM is still running its community services using FirstClass. This is mainly due to the fact that changing abruptly the communicative infrastructure would have too relevant an impact on users’ habits. However, new services are developed gradually adopting open source technologies (e.g., Web applications integrated with FirstClass-based user authentication and forums).

The central element of a FirstClass system are the *forums* (or the so-called “conferences”): discussion areas where information publishing takes place by sending messages. When a community member is interested in a specific forum, she can subscribe to it, that is, put a link to it into her personal home page. New messages sent in the forum are marked with a red flag icon appearing near the forum icon, in order to enable community subscribers to monitor a forum activity simply by logging into the system and scanning through icons on her home page. Messages are then displayed to community members organized in lists, showing the author, the subject and the creation date, but not the body of the message, that can be accessed only by following an appropriate link.

Obviously this approach requires a certain amount of overhead activities for the community member: actually, for being aware of the real relevance of posts and people interacting in the discussion area, she has to log into the system, scan the forums’ icons, open the forums she is interested into, check for new messages, open them in order to access their contents (with no prior indication of their relevance to the discussion or to her interests). Moreover, at the moment, a notification mechanism via email for subscribed forums still does not exist.

From a more technical point of view, forum and messages are managed in different ways: forums can be nested, forming a tree structure, while messages are all at the

same level, but can be grouped into threads. Forums can be both public or private: in particular, in public ones messages can be posted only by authenticated community members, but can be read by anyone visiting the CN Website.

Forum and message list are rendered via Web as dynamic pages generated through templates coded as a mix of HTML (for the fixed part of the page: headers, frame disposition, etc.) and Server-Side-Include-like (SSI) scripting language (for the variable part: the message list, the forum list and so on). These templates can be accessed and customized by the FirstClass administrator. However, customization cannot alter:

- the order in which templates are picked-up to compose the whole page;
- the availability of objects in the rendering phase (e.g. the body of a message is unavailable – due to internal server processing constraints - during the message list composition).

Even though the FirstClass customization features could be exploited to produce an XML/RSS version of the message list, two main drawbacks remained:

- the list of items could not contain the description of the news (i.e. the body of the message), due to the above mentioned processing constraints;
- every forums can have its own RSS feed, but there is no way to integrate in one single RSS feed contents derived from two or more different forums.

These two drawbacks imply that the RSS feeds would reproduce in a “local version” exactly the same structure of the online system, hence also with the same overheads we have described before for the community member accessing the contents.

4 The Dynamo Extractor

The Dynamo Project³ (Bossa, 2005, Bossa et al., 2006) addresses data extraction and channeling over legacy Web sites in plain HTML. Dynamo is intended to benefit two types of users. First, webmasters may employ it to manage the creation of RSS feeds, thus avoiding to do it by hand or by means of proprietary software. Second, users, i.e., consumers of feeds, may use it to overcome limitations such as i) old feeds may not be consulted and usually are deleted from servers and ii) traditional HTML servers cannot execute advanced queries *directly*.

On the contrary, with Dynamo it becomes possible to:

- automatically and dynamically generate RSS feeds starting from HTML Web pages;
- store feeds in chronological order;
- query and aggregate them thanks to Web Services (WS) acting as agents.

It is important to stress that these results were obtained with a lightweight pull algorithm for retrieving HTML documents by Web servers, thus minimizing the required Web traffic for the updates of news sources (Bossa et al., 2006).

³ Dynamo is an open-source project available, under GPL, from <http://dynamo.dynalias.org/>

HTML documents contain a mixture of information to be published, i.e., meaningful to humans, and of directives, in the form of tags, that are meaningful to the browsers and determine the appearance on the screen. Moreover, since the HTML format is designed for visualization purposes only, its tags do not allow sophisticated machine processing of the information contained therein.

Among other things, one factor that may prevent the spread of the Semantic Web is the complexity of extracting, from existing, heterogeneous HTML documents machine-readable information. Although the Dynamo project addresses only a fraction of the Semantic Web vision, our management of HTML documents needs some technique to locate and extract some valuable and meaningful content. Therefore, a set of annotations, in form of meta-tags, were defined; they are inserted inside HTML in order to highlight informational content that is essential for the creation of a RSS feed. In our application, meta-tags are used as annotations, to describe and mark all interesting information, in order to help in the extraction and so-called *XML-ization* phases. Notice that with pages that are dynamically generated out of some template (which is the case with practically all on-line fora) Dynamo annotation is done, manually but only once and for all, over the page template.

Once HTML documents are processed by Dynamo, annotated semantic structures are extracted and organized into a simple XML format to be stored and used as a starting point for document querying and transformation. The structure of the XML output resembles the structure of meta-tags previously defined and the RSS XML structure, in order to facilitate transformations from the former to the latter.

4.1 Dynamo: Structure and Underlying Technical Choices

In order to allow a full transparency with respect to the action of scripting languages (for example, ASP, PHP, JSP) that produce dynamic [X]HTML pages, Dynamo meta-tags are enclosed inside HTML comment tags, which implies that both Web browsers and scripting language interpreters simply ignore them.

Another important element is the possibility of querying a Dynamo database through Web services which, for sake of simplicity, have been designed for the REST (Fielding, 2000) style of interaction (see further).

Dynamo is a Java application with a modular structure, for maximizing the flexibility and the extensibility of configuration. Three levels can be distinguished:

- the *Physical Data Storage Level*. It is the lowermost level, which stores resources, and provides a means for retrieving and querying them. It can be implemented over either relational or XML database Management Systems (DBMSs);
- the *Core Level*, which holds the core part of the entire architecture, including the software components that implement the logic of information management and processing. Each component can be implemented using different strategies or algorithms, and plugged into the system without affecting other components, i.e., by simply tuning the application configuration files;
- the *Service Level*, which is the highest level, interacting with Web clients by means of REST Web services.

A more detailed description follows.

The Physical Data Storage Level can be implemented using various techniques. Currently, Dynamo uses the open-source native XML database Exists (Bourret). This choice allows to store and manage XML documents produced by the Wrapper software component in their native format, and to use the powerful XQuery language (Xquery, 2005) for advanced content querying and aggregation. The native XML database is organized as a set of collections of XML resources, where the nesting of collections is allowed. In our application, we store XML resources as provided by the Wrapper software component, one collection for each resource. Each collection holds the various chronological versions of the resource: so, each collection effectively contains the history of the resource, all its informational content and a changelog.

When a new resource has to be stored, the DataManager component performs checks to avoid duplicate resources. Two resources are considered different when their informational content changes. More precisely, they are different if changes to titles, links or descriptions of the resource channel or items are detected. Once stored, the resource is chronologically archived and ready for later retrieving and querying.

The Core Level consists of several components which define how Dynamo extracts relevant information from Web sources, processes them in order to extract semantic information and finally formats them for clients' consumption:

- the *Poller* monitors changes in a set of HTML sources (defined in a particular configuration file) using a flat polling policy, that is at regular time intervals or, even better, a “smart” polling policy. The latter consists in estimating the time of the next publication of the news on the basis of previously published news;
- the *Retriever*, invoked by the Poller, captures the HTML files and passes them to other components for further computation and storing;
- the *Wrapper*, which extracts the semantically relevant information from HTML files and creates an XML file containing the desired informational content;
- the *DataManager*, a gateway to the Physical Data Storage Level. It takes care of managing information in the form of the new XML documents previously created, storing them and permitting client components to query their contents;
- the *Transformer*, which transforms the XML file into any of the desired RSS formats. It uses suitable XSLT transformations in order to do this;
- the *Engine*, which coordinates all previously described components.

The Service Level lets Web clients access the RSS feeds through the use of REST Web Services.

REST is the acronym of *Representational State Transfer*. It is an architectural style which conceives everything as a resource identified by a URI. In particular, it imposes a restriction about of the URL defining the page info, that, in the REST view, are considered resources. Each resource on the Web, such as a particular part specification file, must have a unique URL (without GET fields after it), that totally represents it.

This allows the client (and all proxy/firewall systems in the middle) to define the next state change only by inspect the URL of current available forwarding links (for the proxy/firewall systems, only by inspect the header of the HTTP request).

With respect to the well-known SOAP architecture (SOAP, 2003), in REST we never access a method on a service, but rather a resource on the Web, directly using the standard HTTP protocol and its methods, i.e., GET, POST, PUT and DELETE.

This feature of REST allows greater simplicity and maximum inter-operability with any Web client, either *thick*, like a desktop application, or *thin*, like a Web browser.

5 Applying Dynamo to the RCM Community

As shown above, the template generation model adopted by the FirstClass system, although widely flexible, cannot be used directly for generating a *useful* RSS index file. However, the template model easily permits to put the special mark tags inside Web pages and to carefully choose where to place tags inside the text. This option, however, is an all-or-nothing type: either all generated pages are marked or none will. As a results, experimentations can take place according to an incremental schema, since single subgroups of forums can be processed separately, selecting them accordingly whichever criterion we feel adequate for testing a specific aspect or functionality of Dynamo (e.g. semantic similarity, frequency of posts, presence of attachments/pictures, etc.). Presently we have tested the system for collecting the *title* and the *URL* of the forum, that are used to populate the channel description part of the RSS feed *index*, and the *author*, *title*, *date* and *body* of the messages posted in the forum, that are needed for filling the *item* part of the feed. The tags used to guide Dynamo during the extraction are described in the following table:

TAG	Description
<!-- <channel:title> --> <!-- </channel:title> -->	the forum title
<!-- <item:author> --> <!-- </item:author> -->	marks the sender of the message
<!-- <item:link index="n"> --> <!-- </item:link> -->	marks the link of the <i>n</i> -th message
<!-- <item:title index="n"> --> <!-- </item:title> -->	marks the subject of the <i>n</i> -th message
<!-- <item:extension localName="date"> --> <!-- </item:extension> -->	marks the date of the message

Thanks to the insertion of the tags, the Dynamo servlet can collect the marked-up information by periodically polling the RCM forums⁴, and store the parsed information into the Exists database. All the RSS generation process, polling, parsing and deploy RSS feeds, are performed by a distinct host respect to that used for RCM framework, running a Tomcat server for the Dynamo engine and a standalone instance on the *Exist* open-source, native-XML DBMS.

6 Open Problems and Future Work

This article described the deployment of the Dynamo data extraction tool to the RCM community Web site.

⁴ Currently the polling period is set to one minute. See Bossa et al. (2006) for a discussion of the tuning of polling policies.

The deployment of Dynamo-based RSS services to RCM forums is very recent and meaningful statistical data over adoption by the community is not yet available. Hence, the impact of the introduction of Dynamo RSSs over the community could not be assessed at this stage

The most visible result of this partnership is the solution of the *legacy barrier* that prevented RCM, locked into proprietary and perhaps a little outdated software, to offer to its users the now-standard RSS feed service (this result could be also beneficial to other communities who are using FirstClass or similar software products). Also, the availability of RSSs may somewhat facilitate (and simplify) community research over RCM.

Less visible but very interesting, in view of future developments, is the possibility to conceive and deploy sophisticated aggregation policies for the contents extracted from the community's forums. Since all RCM forums are now marked up with Dynamo meta-tags, users may effectively customize their feed channels to suit their interests and perception associated to each forum. To do so, we are developing a Dynamo customization tool to be offered on the Web to RCM users who want to decide a personal *information mix* over the several forums she may want to consult at once.

Acknowledgements

Sergio Bossa conceived and implemented the first version of Dynamo.

References

- Armstrong, A.G. and J.III Hagel (1998). Net Gain – creare nuovi mercati con Internet. Etas (in Italian).
- Atom Enabled (2005). Atom Syndication Format (RFC4287). <http://www.atomenabled.org/developers/syndication>, <http://tools.ietf.org/html/4287>.
- Baumgartner R. et al. (2005). Web Data Extraction for Business Intelligence: the LiXto Approach. Proc. of BTW Workshop.
- Bossa, S. (2005). Gradation Project in Informatics. University of Messina (in Italian).
- Bossa, S. Fiumara, G. and Provetti, A. (2006). A Lightweight Architecture for RSS Polling of Arbitrary Web sources. Proc. of WOA conference. Available from <http://mag.dsi.unimi.it/>
- Bourret, R. P., XML and Databases. <http://www.rpbourret.com/xml/XMLAndDatabases.htm>
- Cova, B. (2003). "Il marketing tribale: legame, comunità, autenticità come valori del Marketing Mediterraneo.." Il Sole 24 ORE (in Italian).
- De Cindio, F., Gentile, O., Grew, P. and Redolfi, D. (2003). Community Networks: Rules of Behavior and Social Structure in The Information Society, special Issue "ICTs and Community Networking", Sawhney H. (ed.), vol. 19, n. 5, pp. 395-406, November-December 2003.
- De Cindio, F. and Ripamonti, L.A. (2006). Natures and Roles for Community Networks in the Information Society. Invited paper in P.Day (Ed.), AI & Society special issue on "Community Informatics."
- Fielding, R. T. (2000). Architectural Styles and the Design of Network-based Software
- Gottlob, G. and Koch, C. (2004). Monadic Datalog and the Expressive Power of Languages for Web Information Extraction. Journal of the ACM 51.

- Kozinets, R.V. (1999). E-tribalized marketing? The strategic implications of virtual communities of consumption. *European Management Journal*, Vol.17, N.3, pp.252-264.
- Lipnack, J. and Stamps, J. (2000). *Virtual Teams: People Working Across Boundaries with Technology*. Wiley.
- Polanyi, M. (1967). *The Tacit Dimension*. Routledge and Kegan Paul, London.
- Simon, H. A. (1971). *Designing Organizations for an Information-Rich World*. In Martin Greenberger, ed., *Computers, Communication, and the Public Interest*, The Johns Hopkins Press, Baltimore, MD, ISBN 080181135X.
- SOAP (2003), SOAP v. 1.2. Par 0: <http://www.w3.org/TR/2003/REC-soap12-part0-20030624/>
- UserLand (2005). RSS 2.0 Specifications. <http://blogs.law.harvard.edu/tech/rss>
- Wenger, E., R. McDermott, and W.M. Snyder (2002). *Cultivating communities of practice - A guide to managing knowledge*. Harvard Business School Press, Boston, MA, USA.
- Xquery (2005), XQuery 1.0 : An XML Query Language, <http://www.w3c.org/TR/xquery>

Incorporating Indigenous World Views in Community Informatics

Larry Stillman¹ and Barbara Craig²

¹ Monash University, Australia

² Victoria University of Wellington, Aotearoa/New Zealand
{larrys@fastmail.fm, Barbara.Craig@vuw.ac.nz}

Abstract. This paper aims to provoke further theorising and action by the Community Informatics community about working with Indigenous communities. In particular, we present research undertaken with the Indigenous Maori and *Pakeha* (European) community in Aotearoa/New Zealand as a case study to learn from. Maori are of interest because of their engagement with, and speaking out, about ICTs. We suggest that particular attention needs to be paid from an ethical perspective in working with diversity in order that research and action are undertaken that benefits both the researcher and participant community. Community Informatics would benefit from more attention to articulating its assumptions about the nature of research and action with cultural diversity in its role as a bridge between diverse communities and the design and implementation of Information and Communication Technologies.

Keywords: Community technology; community informatics; Maoris and technology; *Kaupapa Maori* research (Indigenous Knowledge); New Zealand and technology.

1 Introduction

In June 2005, a number of community informatics researchers and practitioners came together in the UK to discuss qualitative research issues in Community Informatics (CI) ¹. A strong concern was expressed about the power imbalance between the researcher and the researched, including work with minority and Indigenous communities. It was felt that such concerns needed further, in-depth exploration, in order that CI develop a more sophisticated and ethical approach to action and research with non-Western communities, given the history of exploitation of Indigenous people through all sorts of research and practice projects that have frequently benefited the researcher far more than the researched [1].

This problem is further complicated by Western researchers inadequately representing the experience and worldview of the 'other' in their research accounts, an experience well-known to many Indigenous peoples around the world. CI cannot stand outside of this historical experience. The history of iniquitous relationships

¹ <http://kmi.open.ac.uk/events/ci2005/pmwiki.php/Together/Summary#theme1> (Accessed: 1 October, 2005).

means that we should seek a form of CI practice that does not imply dependency *upon*, or *patronisation* because of researcher or practitioner expertise, but instead, *partnership with*. It is for Indigenous people to decide what they wish to do with ICTs.

But how we go about this is not well-understood. Research about human actors in social-technical systems is well-advanced, for example, in the corporate world. Orlikowski has convincingly shown that the take-up of ICTs is closely linked to cultures and sub-cultures in corporations, and that these have powerful agency [2, 3]. Similar studies of Indigenous communities or lesser known forms of social organisation such as the welfare sector in developed countries are not so common [4].

Similar concerns have been raised by Salvador and Sherry in an account of their ethnographic work for Intel in South America and elsewhere. They spoke of the need to have a deeper understanding and ‘enliven the lived experience’ of the ‘local’ in the intersection between ‘people and places’ in order that technology design has real meaning in local contexts [5]. In addition, a draft Ethics Statement for CI including research with Indigenous communities is being written². Such statements should include a specific reference to the need for developing cultural safety and cultural competence. Cultural safety and cultural respect are more than recognition of the ‘other’ on the part of the expert or practitioner: they actively partner knowledge and skills held by the different parties (such as the researcher and members of an Indigenous people)[6]. This paper is intended as a contribution to partnerships with Indigenous communities as they relate to localised social-technical opportunities³.

2 The World of the New Zealand Maori

New Zealand’s geographic isolation, small population (just over 4 million people), and ecological and economic fragility have prompted considerable interest in ICTs from its national government as a means to use ICTs for building a better society and to more effectively connect it to the rest of the world [7, 8]. *Aotearoa* is one of the last places on earth to be settled by people. Maori arrived there from eastern Polynesia about 800 years ago, and European colonisation only took hold in the latter half of the eighteenth century. New Zealanders are thus conscious of their origins in ways that are not found elsewhere in the world and for many *Pakeha* (European origin) New Zealanders, and immigrants from other cultures, their life in these islands in the Pacific is increasingly lived as a multicultural encounter with the values and heritage of the Maori, the *tangata whenua*⁴, the ‘people of the land’.

² See (Draft) Code of Ethics for Community Informatics Researchers, <http://vancouvercommunity.net/lists/arc/ciresearchers/2006-07/msg00024.html>, 27 July 2006.

³ We do not speak with a definitive or authentic or authoritative voice for Maori. Furthermore, this paper betrays our ideals by not being the product of a much more collaborative process with our interviews and fieldwork participants. Our paper should be therefore regarded as the introduction to, rather than end-point of ongoing research and action and a step in our and the CI community’s learning.

⁴ See http://en.wikipedia.org/wiki/Tangata_Whenua

Te Tiriti, the Treaty of Waitangi⁵ (1840), in which particular rights were granted, serves as a tool ‘by which we can measure the benefits, and make use of existing structures within Maori societies [9, 10]. However, the Treaty guarantees were never upheld and *Pakeha* dominance has persisted in many arenas of social life, including, of special interest here, the conduct of community-based research and action. This has led to the development of a particular set of protocols for research called *Kaupapa Maori* research (Indigenous Knowledge). This challenges traditional Western research practices in its stance of being collectivist and participatory, and in which the subjects of research and action have a strong and determining voice [11].

3 The Tuhoe Tribe as a Case Study⁶

Maori are users of ICTs, even though lower direct ownership levels of computers are more a problem of cost and low socio-economic status than attitude to technology [12]. In the recent New Zealand government statement about community connectivity, it is made clear that the term ‘community’ takes on special meanings in Aotearoa and that technology has a powerful role in maintaining traditional social forms, as well as forging new relationships between *Pakeha* (European inhabitants) and Maori and other cultural groups:

ICT can enhance our sense of identity and connection to a particular place or group. It can extend services to isolated communities or those excluded from full participation in the life of the community. It can enable people to become more involved in democratic processes and decision-making at all levels. The government recognises the vital role that community, voluntary, and Māori organisations and *iwi* play in New Zealand society. ‘Community’ means more than geographic communities. The term includes traditional associations such as *whānau* and *hapū*, ethnicity or occupation, and virtual communities of interest or practice [7: 33].

Recent New Zealand government policies include extending broadband through experimental technologies into remote rural communities and linking remote Maori boarding schools through video so that all members of an *iwi* can be connected (whether living in urban settings or on tribal marae) and be involved in decision-making about their futures. Funding for remote and under-served community broadband is available through the Digital Strategy’s Broadband Challenge Fund. Recent funding includes the Tuhoe community network.

Tuhoe are a Maori *iwi* (or tribe), whose traditional land (Tuhoe *rohe*) is in the eastern part of New Zealand’s North Island. Tuhoe is a confederation of 25 *hapu* (extended families or subtribes) and 40 marae (formal meeting spaces) linked by

⁵ http://en.wikipedia.org/wiki/Treaty_of_Waitangi

⁶ This research was conducted by Craig and others as part of the 2020 Trust Computers in Homes project which gave 200 Tuhoe *whanau* (family) computers, training and internet connections as a first step in the digitising of *mataruanga Tuhoe*, that is, Maori research and resources including physical, oral, literary, artistic, and other means (see <http://www.tuhoematauranga.org.nz/>).

whakapapa (geneology). They have a strong Maori identity, including use of the Maori language. At the 2001 NZ census, 25,000 lived in the tribal areas, with another 25,000 widely dispersed globally and throughout other New Zealand communities. Traditional communities are sprinkled throughout a remote, rural mountainous area with very poor infrastructure and access to services such as telephone and electricity. Yet ICTs play an important role. Interviews with residents show how Webcams, MSN and chatrooms are all used for communication.

My uncle wants to set up a chatroom with all my cousins I haven't seen or heard from since they were like babies (sic). I am starting to learn how to do that.

These communities are economically depressed and geographically isolated, but with a very strong local identity and culture. ICTs in these communities are tools for connecting tribal members so that all Tuhoe can be engaged in local decision-making and knowledge can be better shared. ICTs are also powerful tools for maintaining identity, culture and language and to place Tuhoe firmly on the digital landscape. Tuhoe seek alliances with various Pakeha agencies such as universities and government to appropriate the benefits of global communications for their peoples collectively. Through such partnerships they have digitised their collective history, installed WIFI across the valleys and put videoconferencing in the schools and computers with internet access into the homes. Interviews with families with home computers show this collective and reciprocal arrangement rather than individual ownership of ICTS in these communities, in contrast to the individualism and private arrangements that are more familiar to *Pakeha*.

The whole community comes over to use it. They play cards. They found this computer game and they put it on and gradually the game was killing everything on the machine.

There are just three of us at home. I have got a few *whanau* [family] boys and stuff that come around. Usually while I am at work they will come over during the day and thrash it [the computer] to bits.

Intergenerational learning is another collective way of interacting with technology in these communities. Grandparents traditionally spend a lot of time with their *mokopuna*, their grandchildren. This younger generation have lots of access to ICTs in school and grandparents depend on their *mokopunas'* help after school with using the internet for finding information or using their email. An interview with a school principal suggested that intergenerational learning with ICTs could help foster the Maori language and culture in the young generation at the same time as bringing tribal elders up to speed with new technologies.

We had a couple of grandparents online but they gave up because their helpers – the young ones- were moving too fast. If the language of communication was Maori that will slow them down. Most of our children are not bad at both languages but our young parents have a bit of a problem and come here to school for Maori classes....our grandparents are best with Maori.

4 Other Perspectives

For a researcher or practitioner concerned to work effectively with Maori, kin relationship and connection, what is called *whanaungatanga*, are at the core of any process of engagement. Bishop, a proponent of *Kaupapa Maori research*, suggests that establishing a research group as if it were an extended family is one form of utilising what he refers to as the ‘treasures of the ancestors’, the collective wisdom of the ages, that guides and monitors everyday practice [11: 128]. This approach to research is about re-creating the infrastructure of reciprocity and relationship in a culturally appropriate way.

While the utilisation of such a framework, and particularly, the process of engagement with community could be dismissed as distracting and time-consuming background noise that inhibits ‘efficient’ Western-style research, they are clearly real and relevant factors with strong agency and with which the researcher must engage. In fact, the process of reflexive practice and empathetic engagement with the research partner/s (called by Giddens the double-hermeneutic [13]), is integral to qualitative participatory research but here it located within the Maori world view, not that of the researcher.

This collectivist approach is not confined to academics, but is common practice for Maori in their communities in New Zealand. In other conversations conducted for a set of research interviews in New Zealand⁷, both Maori and *Pakeha* described very particular world views. Maori concepts and words were used to describe the world view, affiliations, genealogical and historical and cultural connectedness that people have with their communities and how ICTs are part of that world. In particular, the importance of genealogy and collective, rather than individual affiliation and identity were emphasised.

Thus, a Maori interviewee with a high level of online skill said that:

You know, as Maori we all have responsibilities as well, so we all have a responsibility to our families, our immediate families, but to the wider community. So if you’re educated, your responsibility and the expectation is that you will contribute to your lives, to your land, to the development of your hapu, yes your tribe. Now that’s the expectation ...we’re constantly [in demand], because we don’t have a lot of people that are skilled in those provinces, those of us who have those skills are expected to share them.

This orientation towards a community, rather than individual use of technology is also addressed by comments from another interviewee. *Whanaungatanga* is the term used by one *Pakeha* interviewee in discussing her technology work with Maori people:

It’s tied up with trust, it’s a safe place for them to come to, it’s familiar, somewhere where the learning, particularly out in the Maori rural areas, because the learning operates on what we call *whanaungatanga*, *whanau* is family...It is the act of doing it, it’s the inclusive family relationship that

⁷ These were conducted as part of Stillman’s PhD research.

goes with learning, so that it's not about the individual student coming along to class, to learn, bugger you, I'm just going to do what I want to do, it's all about learning as a group, and the success of the school is the success of the group.

And another person said:

[P]eople are taught about IT in the individual sense, so when they talk about user needs, they'll go and study how individuals work. But say if we are going to have a plan to provide, I don't know, broadband to a Maori, it's not about individuals is it? It's about how does the collective get their service? And I mean talking to them, obviously all of them, or developing tools that are meaningful for a group, and not just an individual, you see.

Another, non-Maori interviewee, experienced in community development work with particular disadvantaged communities, said that:

Tertiary [i.e. university] education has become about bums on seats, there is a need to use Maori terminology. *Awhi* is the word, *awhi*, is to nurture, when you *awhi* someone, you nurture and mentor them along...there's a lot of *awhi*, *awhi*, in that process of Maori learning, and why [Technology] Project has been embraced a lot, by the Maori community and schools, because as much as possible we go in alongside them, and work with how they are, and how they want to work, rather than imposing something from the top down, you poor people, this will be good for you. You'd get short shrift out the door.

At least one Maori writer with links into both academia and community action has articulated a particular, Maori theory of technology, taking account of cultural issues and the history of exploitation, assimilation, and racism. Interestingly, her definition of information technology is one that is beyond the technical artefact, but refers to cultural processes and actions as well:

Definitions of information technology need not be limited to those found in information technology journals. Potentially, any means of storing, analysing and disseminating information can be included – even our minds. By ignoring the jargon and focusing on this idea, it is clear that Maori concepts such as *matauranga* and *hinengaro* can encapsulate (and enhance) what we believe about information technology and offer a wider context. *Matauranga* refers to education and intuitive intelligence, and is linked to the divine. *Hinengaro* is the mind, the thinking, knowing, perceiving, remembering, recognising, feeling, abstracting, generalising, sensing, responding and reacting ...In this light, Maori knowledge informs us about why Maori might be highly motivated to take up information technology and why concepts of information technology, as its industry sees it, are not only accessible to Maori but even simplistic [14: 466].

This assertion of an affirmative, localised, and Indigenous response to technology is a response to colonisation. As a *Pakeha* interviewee put it in explaining the different views of technology which he has encountered in the Maori community:

[People say] “Well, we were colonised once, and there is a great possibility that we are going through a digital colonisation phase as well”. But they’re saying, during the original colonisation, the white people defined their history. They gave them the literacy to do that. They defined their history for them. And the culture that the white people were coming from had no context to explain the culture of the Maori people.

In response to that loss caused by European domination in the past of their culture and history (and the potential for future loss), Maori concerned about ICTs and their people have increasingly developed their own understanding of how technology can be used for the cultural protection and production.

5 Implications for Community Informatics

Community informatics (CI) is an emergent discipline and practice that draws upon social and technical expertise. According to what is regarded as a consensus position [15] in Wikipedia:

Community Informatics... refers to an emerging set of principles and practices concerned with the use of Information and Communications Technologies (ICTs) ... for enabling the achievement of collaboratively determined community goals; and for invigorating and empowering communities in relation to their larger social, economic, cultural and political environments. [16]

What Maori bring to the table is a particularly powerful Indigenous articulation of a philosophy of collaborative participatory research [17, 18] to take up the challenge of collaboration presented above. Indigenous culture has been too often regarded as primitive, easily acquired, and secondary to the expert knowledge (including technical knowledge) held by the Western-trained researcher or practitioner. As a consequence, the history of action and research (even of the most well-intentioned sort) with Indigenous communities has all too frequently been bound up with issues of colonising epistemologies, unequal benefits of research, and deterministic cultural, political, and economic agendas in favour of the researcher, rather than the ‘researched’ [11]. The same comment is relevant to socio-technical agendas, even if well-intentioned.

Thus, a real concern by some Maori continues to be the ongoing appropriation of ‘Indigenous knowledge, system of classification, technologies and codes of social life’ into electronic networks and their potential misuse [1: 60]. Maori have been particularly concerned about the appropriation of biological and genealogical records for public distribution, and these pose difficult ethical and practical challenges because of the spiritual and cultural significance of such data to them. Others are

concerned about the individualist economic push revealed in ICT policy, rather than a commitment to reciprocity. The challenge is to work with communities to find socio-technical solutions to the proposition that ‘Indigenous peoples are not merely stakeholders in their heritage—they own that heritage and that the right to fully control and if and how research is undertaken on that heritage’ [19: 236].

This is no easy task. We need to be prepared to allow communities to develop research at their own pace and through their own processes of governance, in conjunction with outsiders so that far more equitable power relations are established. Kamira suggests that a key principle which underlies this *kaitiakitanga*, in which there is ‘guardianship, protection, care and vigilance of data about Maori that is collected, stored and accessed’ [9]. Such statements have enormous significance for how ICTs are to be presented, controlled, and put into effective use [20] by such communities.

6 Conclusions

Via a case study of what we have been able to discern about Maori understandings of the place of technology in their social and cultural development—and the way in which knowledge about their community is governed—we have hoped to make clearer the challenge of invigorating and empowering indigenous people on their own terms. What is significant about Maori is that they have begun to articulate a particular theory of ‘collaborative’ research and its location in the family and tribe and this provides a very clear conceptual and practice base for CI researchers and practitioners in New Zealand to work with. It is an example to be considered elsewhere. It is necessary for CI in fact, to re-interpret the very concept of ‘technology’ if it is to work well with such communities. From the perspective of welfare and community research, as well as in the writings of Foucault or Heidegger, ‘technology’ can be re-interpreted to be seen as a network of human processes or practices and techniques involving the use of resources and power, incorporating a body of knowledge and practice which can be complemented by ICTs [21, 22].

A broader definition of technology hearkens back to classical understandings of technology as a culturally-embedded skill and process [23], but it is also alluded to in the comments of Kamira, quoted previously, as relevant to Maori interaction with ICTs [14]. The more specifically ‘technical’ aspects of CI, which draw upon expertise in Information Systems and other disciplines can be considered as part of the basket or mix of social and artefactual technologies that are drawn upon in the development of community technologies such as community networks, specialist knowledge systems, databases and so on, but at the same time, social change and development processes that benefit all parties [24].

The ways in which indigenous peoples like Maori understand the process of research and action should be brought into the discussion of CI so that it develops a richer and more ethical appreciation of different ways of working with other peoples. Additionally, CI needs to embrace a broader idea of what we mean by technology and be more prepared to thoroughly and patiently negotiate and engage in a contestable discourse about what the ‘social’ side of socio-technical solutions means and then work on technical solutions in close and astute partnership for self-determination.

References

1. Smith, L.T., *Decolonizing methodologies : research and indigenous peoples*. 1999, London; New York: Zed Books
2. Orlikowski, W.J., *Using technology and constituting structures: A practice lens for studying technology in organizations*. *Organization Science*, 2000. 11(Jul/Aug): p. 404-428.
3. Orlikowski, W.J., *The Duality of Technology: Rethinking the Concept of Technology in Organizations*. *Organization Science* 1992. 3(3): p. 398-427.
4. Harlow, E. and S.A. Webb, *Information and communication technologies in the welfare services*. 2003, London ; Philadelphia, Pa.: Jessica Kingsley Publishers.
5. Salvador, T. and J. Sherry, *Local learnings: an essay on designing to facilitate effective use of ICTs* *Journal of Community Informatics*, 2004. 1(1): p. 76-83.
6. Australian Indigenous Doctors Association. *An Introduction to cultural competency*. 2004 [cited 2006 1 August]; Available from: http://www.racp.edu.au/hpu/policy/indig_cultural_competence.htm.
7. Government of New Zealand, *The digital strategy: creating our digital future*. 2005, NZ Government. Ministries of Economic Development, Health, Research Science and Technology and Education: Wellington, NZ.
8. Williamson, A., *A Review of New Zealand's Digital Strategy*. *The Journal of Community Informatics*, 2005. 2(1): p. 71-75.
9. Kamira, R., *Kaitiakitanga: Introducing useful Indigenous concepts of governance in the health sector*, in *Information Technology and Indigenous People*, L.E. Dyson, M. Hendriks, and S. Grant, Editors. 2007, Idea Group Inc.: Hershey, PA.
10. King, M., *The Penguin History of New Zealand*. 2003, Auckland, NZ: Penguin.
11. Bishop, R., *Freeing Ourselves from Neo-colonial Domination in Research: A Kaupapa Māori Approach to Creating Knowledge*, in *The SAGE handbook of qualitative research*, N.K. Denzin and Y.S. Lincoln, Editors. 2005, Sage Publications: Thousand Oaks.
12. Parker, B., *Maori access to information technology*. *The Electronic Library*, 2003. 21(5): p. 456-460.
13. Giddens, A., *The constitution of society : outline of the theory of structuration*. 1984, Berkeley: University of California Press.
14. Kamira, R., *Te Mata o te Tai – the edge of the tide: rising capacity in information technology of Maori in Aotearoa-New Zealand*. *The Electronic Library*, 2003. 21(5): p. 465-475.
15. Gurstein, M., *Personal Communication*. 2006.
16. Wikipedia. *Community Informatics*. 2006 [cited 2006 17 April]; Available from: http://en.wikipedia.org/wiki/Community_informatics.
17. Stillman, L., *Participatory Action Research for Electronic Community Networking Projects*. *Community Development: Journal of the Community Development Society*, 2005. 36(1): p. 77-92.
18. Stoecker, R., *Research methods for community change : a project-based approach*. 2005, Thousand Oaks: Sage Publications.
19. Niven, I.J. and L. Russell, *Appropriated pasts: indigenous peoples and the colonial culture of archaeology*. 2005, Lanham: Altamira.
20. Gurstein, M., *Effective use: a community informatics strategy beyond the digital divide*. *First Monday*, 2003. 8(12).
21. Foucault, M., *Technologies of the Self: A Seminar with Michel Foucault*, ed. L.H. Martin. 1988, London: Tavistock.

22. Heidegger, M., *The question concerning technology*, in *The Question Concerning Technology and Other Essays*, W. Lovitt, Editor. 1977, Garland Publishing, Inc.: New York. p. 3-35.
23. Bell, D., *The winding passage : essays and sociological journeys, 1960-1980*. 1980, Cambridge, Mass.: Abt Books.
24. Charmaz, K., *Grounded theory in the 21st century: applications for advancing social justice studies*, in *The SAGE handbook of qualitative research*, N.K. Denzin and Y.S. Lincoln, Editors. 2005, Sage Publications: Thousand Oaks.

Towards a Theory of Online Social Rights

Brian Whitworth¹, Aldo de Moor², and Tong Liu¹

¹ Massey University (Albany), Auckland, New Zealand

b.whitworth@massey.ac.nz, T.Liu@massey.ac.nz

² STARLab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
ademoor@vub.ac.be

Abstract. Legitimacy, defined as fairness plus public good, is a proposed necessary online and physical community requirement. As Fukuyama notes, legitimate societies tend to prosper, while others ignore legitimacy at their peril. Online communities are social-technical systems (STS), built upon social requirements as well as technical ones like bandwidth. As technical problems are increasingly solved, social problems like spam rise in relevance. If software can do almost anything in cyberspace, there is still the challenge of what should it do? Guidelines are needed. We suggest that online communities could decide information rights as communities decide physical action rights, by a legitimacy analysis. This requires a framework to specify social rights in information terms. To bridge the social-technical gap, between what communities want and technology does, rights must be translated into information terms. Our framework has four elements: information actors (people, groups, agents), information objects (persona, containers, items, comments, mail, votes), information methods (create, delete, edit, view, move, display, transfer and delegate), and the information context. The conclusions apply to any social-technical community, and we apply the framework to the case of Wikipedia.

Introduction

It has been proposed that legitimacy is necessary for social productivity whether based on electronic or physical media [10], where legitimacy is defined as fairness plus social good. In this view email, chat, bulletin boards and groupware are *social-technical systems* (STS), i.e. social systems that overlay technical ones, where “technical” includes both software and hardware aspects. IS theory suggests information systems have different levels: Grudin suggests three: hardware, software and cognitive [4], Kuutti adds a work processes level [5], and Alter suggests hardware, software, people, and business processes [2]. Table 1 shows four STS levels, each a “view” of the same system, not different systems. Each level “emerges” from the previous, as information/data derives from mechanics, cognitive meaning has an information base, and community norms arise from human cognitions [11]. Information here is used as Shannon and Weaver originally defined the term [9], and equates to what business calls “data”. Higher levels assume lower levels, so a level failure implies failure at levels above it, e.g. if hardware fails, software does too, as does the user interface, but a program can fail at software level but still function as

hardware. Why not reduce everything to hardware? Describing computing by chip and line events is as inefficient as describing World War II in terms of atoms and electrons. Higher “holistic” levels increase performance, so communities can generate enormous productivity. STS design must reflect social requirements, lest online “civilization” become a stone-age culture built on space-age technology.

Table 1. Information system levels

Level	Examples	Error	Discipline
Social/ Cultural	Norms, culture, laws, sanctions, social roles	Unfairness	Sociology
Personal/ Cognitive	Semantics, attitudes, beliefs, opinions, ideas, ethics	Misunderstanding	Psychology
Information/ Data	Software, data, bandwidth, memory, processing	Infinite Loop	Computing
Mechanical/ Physical	Hardware, computer, telephone, FAX, voltages, heat	Overheating	Engineering

Translating acceptable principles of social interaction into STS specifications can bridge the social-technical gap, between what communities want and what technology does. STS designers must know what to do in information terms, based on legitimacy requirements, expressed as “rights”. Just as a physical community uses ownership to express physical rights, so an online community can use information ownership to express information rights [10]. Legitimate rights like freedom increase social productivity but can do so at the cost of social stability. This social problem has not one but many “solutions”, e.g. there are forms of democracy. Yet while social issues have no perfect answer, not all social forms have equal outcomes, as in general more legitimacy is better than less. We propose a general framework to specify online “rights”, as these are the design building blocks of a healthy online community. We hope others find our framework useful to build upon, modify or even contradict.

Criteria

For a software designer to implement social rules they must be:

1. *Complete*: The computer knows what to do in every case.
2. *Consistent*: The rules are easy to program and apply to new cases.

An online community in contrast desires interaction rules that are legitimate:

1. *Fair*: That the STS is impartial to individual actors, treating them as it were from behind a veil [8]. Justice is blind to individuals, i.e. does not favor specific actors. The social goal is social productivity, not personal gain (which is corruption).
2. *Socially beneficial*: A legitimate STS improves the public good. People tend to cooperate with legitimate communities and oppose illegitimate ones.

Rather than argue what past communities agree, we assume the following:

- I. Social entities should be accountable for their acts to the community, which may apply sanctions like banishment.
- II. To be accountable, social entities must be identified (not anonymous) to that community. Privacy gives the right to be anonymous *to others*, not to the community itself. A society can record data on birth, marriage and death etc, as necessary to identify its members.

Specification

How software architecture allocates information rights defines the social options of a virtual community [6]. We now explore STS rights given the earlier criteria of completeness and consistency (for the computer) and legitimacy (for the community).

STS Primitives

We define the primitives of an STS as social actors, objects and methods (Table 2).

Social Actors. Social actors are people or groups who are accountable to society for their acts. We use “actors” rather than “users” to stress their ability to be accountable. Social agents are actors who act on behalf of another social entity, and can be people or automata, e.g. application installation programs are automated agents for a software company. Independently acting automated entities, however intelligent, are actors but not social actors. They are not accountable, and have no “self” to feel social sanctions. The STS objects that represent social actors can be called *persona*. A persona can be an online “handle”, not a real name, but must be unique. As physical identities make people accountable in that world, persona identities make people accountable in virtual worlds, e.g. one can be banished from an online game. A persona represents when a person “exists” or is active within the STS. When a person “logs on” with userid and password, the STS then sets an active session for them until they “leave”. We call personas “people”, though they just represent them. The basic social actors in Table 2 are people, groups and agents.

Information Objects. Since an STS is an IS, it is an information object (O), as are the objects within it. It can contain *items* (I) whose main purpose is to convey meaning, defined as the cognitive processing evoked in people. If an item’s meaning is dependent upon another source item, it is a *comment item* (IC). Items transmitted between people communicating are *mail items* (IM), and items whose meaning is only choice information can be called *votes* (IV). A *container* (C) is a complex object that can contain other objects, like an item list. Objects exist within containers, and the STS environment is the first C. If a container is destroyed any objects within it are also destroyed, as the existence of O’s in C depend on C existing. A C may contain another C1, in which case C1’s objects are also part of C. The core information objects proposed are persona, containers, items, comments, mail and votes (Table 2).

Table 2. Social-Technical System Components

Social Actors	Objects	Social Methods
<i>People</i> (exist outside the STS)	<i>Persona</i> (represent people)	<i>Create/Delete/Undelete</i>
<i>Groups</i> (composed of people)	<i>Containers</i> (contain objects)	<i>Edit/Revert</i>
<i>Agents</i> (for people/groups)	<i>Items</i> (convey meaning)	<i>Archive/Unarchive</i>
	<i>Comments</i> (dependent meaning)	<i>View/Hide</i>
	<i>Mail</i> (transmit meaning)	<i>Move/Undo</i>
	<i>Votes</i> (choice meaning)	<i>Display/Reject</i>
		<i>Join/Resign</i>
		<i>Include/Exclude</i>
	<i>Rights</i>	<i>Transfer</i>
		<i>Delegate/Undelegate</i>

Social Methods. The actions social actors can apply to information objects include create, delete and edit, and most have an inverse. Some involve two objects, e.g. move changes an object’s container, and can be enter or exit.

STS Rights

Using the components of Table 2, one can define a “right” as follows:

$$\text{Right} = \mathbf{R} (\text{Actor}_i, \text{Object}_i, \text{Method}_i), \text{ or } \mathbf{R}_i = \mathbf{R}(A_i, O_i, M_i)$$

Roles, such as “owner”, are sets of rights, e.g. to "own" an object is the right to all actions on or with that object:

$$\text{Right}_{\text{Owner}} = \mathbf{R}(\text{Actor}_{\text{Owner}}, \text{Object}_{\text{Owned}}, \text{Method}_{\text{All}})$$

Rights Errors. A rights “error” occurs when a party allocated a valid right is unable to exercise it, e.g. when multiple actors have same object rights one party’s rights can deny another’s. If many actors can delete an item, if one person genuinely deletes it then it is gone, so the others have lost their choice to delete or not. The first actor abrogates the rights of others. A simple way to avoid rights errors is for one person to own everything (dictatorship), but this is not fair. In contrast, that everyone owns everything (anarchy) is fair, but invites rights errors and conflict. The middle path between anarchy and dictatorship, that most modern societies pursue, seems based on distributing ownership, and that is the approach we take here. Multiple ownership is complex, as people can act severally (where any can act for all), e.g. husband and wife who trust each other, or act jointly (where all must consent to act), or act democratically (where a majority prevail), or a combination, e.g. democratically elect a leader to act for the group. Ownership can also be passed back and forth, as in joint document writing.

Transfer. Rights built from Table 2’s actors, methods and objects are also information objects in themselves, and shown as second tier objects, subject to meta-actions like transfer and delegation. Transfer changes an object's owner. A right that

incurs no existing information responsibility can be transferred in one step, e.g. a right to view that incurs no viewer obligations may just be given. In contrast, if a transfer incurs information responsibility, both parties must agree in a two-step process:

- a. The owner relinquishes ownership (and may designate the next owner)
- b. The new owner takes up the ownership.

The new owner must agree if the new ownership involves accountability.

Delegation. Transfer gives all action rights to the target, so is non-reversible, but delegation transfers all object rights except the right to change rights, so is reversible, i.e. the owner can take back ownership at any time. The delegatee cannot further transfer ownership, as they have no right to transfer rights, e.g. loaning a book to another gives no right to loan it to a third party. Note that while people can (and do) do this, the issue of whether they should do it is a separate issue.

Object state. An object's state defines the action set that can be performed upon it, with the normal state as "active". In ownership transfer, when owners relinquish ownership they cannot still change the item. In the "given away" state the only action possible is "take ownership", by either the original or new owner(s). When in transfer or delegation, no acts are allowed except take ownership, which returns the object to the "active" state. Another state example is "archived", where no edits can occur but view is still possible, e.g. journal publication is a copyright transfer of ownership followed by an archive state, with viewing but no further edit changes allowed.

Creation Rights

The STS components of Table 2 can define many classes of STS rights, including creation, display, view, comment and group rights, and [10] provides many examples. Space does not permit us to review all these classes here. However, we can briefly discuss creation rights, given the others can be defined analogously.

Object Creation. It is reasonable to assume the initial owner of a created object is its creator, as without them the object would not exist [7]. However in an online setting, where does the right to create come from? An information object comprises various fixed data attributes that must be known before it is created, i.e. it is an instance of a prior general form. If to create an object its general form must be known, the form information must be stored somewhere prior to creation. It cannot be in the object, which is not yet created, so must be in the object's container (or its container, etc, up to the STS). All objects are thus created using information from the object(s) that contain them. Also creating an object also changes the container it becomes part of. Hence it is reasonable to see object creation as an act upon the container the object is created in, which implies the container owner has the right to create objects in it:

Right_{CreateObject} = **R**(Actor_{ContainerOwner}, Object_{Container}, Method_{Create})

Persona Creation. If the STS itself is a container, its owner has the right to create all objects within it. Since persona can act throughout the STS, they seem objects created upon the STS itself, i.e.:

Right_{CreatePersona} = **R**(Actor_{STSOwner}, Object_{STS}, Method_{CreatePersona})

In this case, the STS owner would also own the persona created. However the right to freedom suggests that the person a persona represents should own it.

Right_{Freedom} = **R**(Actor_{PersonRepresented}, Object_{Persona}, Method_{All})

While people normally own what they create (property rights), the right to freedom suggests people should own their online persona selves, which should not be owned by others (slavery). This rights conflict is resolved if the STS owner creates a persona, then transfers its ownership to the person concerned, as many mailing lists do. Systems like Hotmail delegate persona creation, letting entrants self-create persona, as the following right involves no responsibility for existing information:

Right_{CreatePersona} = **R**(Actor_{STSEntrant}, Object_{STS}, Method_{CreatePersona})

If an object owner has all rights to it, this includes the right to destroy it, so one should be able to delete one's persona, e.g. a hotmail-id. However this assumes all transactions are complete, else one could use an online persona to commit say credit card fraud, then "vanish" into thin air by deleting ones persona.

Item Creation. A container owner may delegate their right to create items to people who enter their container, as bulletin board owners let members create items in them. The right can be given freely as no accountability is implied. To create objects within C one may need to "enter" the container, which may be open entry or restricted by a password, equivalent to a door passkey. The delegated right is:

Right_{CreateItem} = **R**(Actor_{ContainerEntrant}, Object_{Container}, Method_{CreateItem})

When an item is added to a list, is it owned by its creator or the list (container) owner, given they are not the same? In the first case, only its creator can delete it, and in the second, only the bulletin board owner can. We propose the initial object owner is always its creator, however the container owner can prevent its display to others (except of course its creator who can always see it) [10].

Creation Constraints. If creation within a container is delegated from the container's owner, the latter can delegate in degrees, i.e. a container may constrain object creation in any way, e.g. a list may require all items be signed. Such constraints should be evident at creation time, giving item creators informed choice. Creation constraints apply only at the moment of creation, so are not retrospective, e.g. if a list allowed anonymous contributions, then required that all contributions be signed, existing anonymous items need not have signatures. The changed creation condition applies only to creations after the change, but if an anonymous item owner wanted to edit it, the new edited item must then be signed (or the edit cancelled).

Creation constraints illustrate a *rights context*, where a right is limited by a contextual right. The delegated right to, say, add an item to a bulletin board can be written:

Right_{CreateItem} = **R**(Actor_{C-Entrant}, Object_{Container}, Method_{Create}, Context_{Constraint})

Wikipedia Ownership: A Rights Analysis

To illustrate a rights analysis we consider Wikipedia, since it is fairly successful, and its philosophy that no-one need own anything is a good test case. Wikipedia holds that all content in Wikipedia is owned by all Wikipedians, apparently currently numbering over 1.9 million. While this is public ownership rather than no ownership, it suggests all rights specifications reduce to a single statement:

$$\mathbf{Right}_{All} = \mathbf{R}(\mathbf{Actor}_{All}, \mathbf{Object}_{All}, \mathbf{Method}_{All})$$

However this utopian specification is not the Wikipedia we actually see today. From its inception, Wikipedia has been under attack by “vandals”, trying to destroy its content integrity with graffiti, pornography, insults or deletions. In response, it has evolved many social rules, which currently involve literally hundreds of pages, detailing rights to edit, to delete, to resign, to join, to create new topics, to revert an item, to change signature etc., e.g. while anyone can edit any item, to create a new item one must first register. Also a social hierarchy has evolved, of stewards, bureaucrats, sysops and other levels including that of Jim Wales, who listens to others but as he notes “...at some ultimate fundamental level, this is how Wikipedia will be run, period”. Even in Wikipedia, the STS owner has absolute rights.

We approach Wikipedia in two ways. First, as a successful online social system, to define generally how Wikipedia allocates various rights, so other applications can implement some or all of them in a different context. Second, a rights analysis may suggest alternative options to those chosen by Wikipedia.

Wikipedian rights. The Wikipedia model has several interesting rights features:

- *Public editing.* A wikipedia creation condition is that the item created is editable by all. When one publishes in a journal one gives them public display rights via a copyright form, i.e. all can view it. Wikipedia simply goes a step further, in that to publish in it, one must give public edit rights.
- *Accountability.* While in Wikipedia anyone can edit anything, it records the IP address, which IP can be banished for community offences.
- *Pseudonymity.* Registering an online pseudonym makes one real world anonymous but still accountable online, as one’s pseudonym reputation affects promotions, and banishment loses reputation gains. All Wikipedia acts are traceable, so all an actor’s acts can be reviewed. If someone vandalizes one item, their other item edits can be checked. Each actor’s “talk page” allows public comments, which they cannot delete or edit.
- *Transparency.* Administrative processes, like steward promotions, are public, i.e. everyone has the right to comment, and everyone can see all comments on position applicants. Final decisions are based on democratic votes.
- *Versions.* After every edit a version copy is kept. Hence nothing on Wikipedia is really deleted, as a “revert” can undo an edit. This reduces rights errors, as no-one ever really loses their rights by permanent deletions.
- *Attribution.* Wikipedia records who made each contribution and so gives unique attribution rights if not unique edit rights.

Wikipedia is an encyclopedia by the people for the people. It engages the power of the community, but must still protect itself against unfair actions like vandalism. In Wikipedia, one “troll” can destroy the good work of many others. Part of that protection is its software base, which implements a rights specification that defines who can do what to what information. Misplaced computer power means a small minority can increasingly damage the majority, e.g. email spam [12].

Wikipedian alternatives. Two alternatives to the Wikipedia rights choices regard ownership of account name and ownership of new item contributions.

Account name. In Wikipedia one’s account name is attached to every online edit. A Wikipedian who initially registers under their real name, like John Doe, then after some edits wishes to change to a pseudonym must ask an administrator to do this, as it affects the Wikipedia database. This creates usurpation problems as one can overwrite an inactive username, i.e. pretend to be a previous contributor. It also means Wikipedian actors have no right to resign, except as permitted. A rights analysis suggests one should own one’s display name entirely. While Wikipedia can create and own unique accountable system ID for each actor, privacy gives actors the right to display themselves to others or not. This suggests two data entities, a “SystemID” known only to, and owned by, the system, and used for community sanctions like banishment, and an “ActorID” or signature, used for public displays. The SystemID never ever changes, so usurpation is impossible. The ActorID is entirely changeable, via an editable profile, so actors need no administrator to change it, and Wikipedia’s change username policy is unnecessary. Wikipedians could genuinely resign, which is not currently allowed without administrative permission, and keep their signature, which no other could then use, or delete it and let another take that signature, making their edits attributed to “Resigned”. The latter illustrates that while physical publishing attributions cannot be changed once done, online publishing authorship allows retrospective reattribution. Changing a Wikipedia signature from Rising Devil to Fallen Angel, gives the system two options. It can retrospectively change all your past edits to the Fallen Angel signature, or it can leave them as Rising Devil but allocate any new edits to Fallen Angel. In the latter case, your signature has effectively two versions arising from your edit of it.

Ownership choice. While Wikipedia favors public ownership it still supports copyright, perhaps because if Wikipedia expects members to follow its rules, it would be inconsistent for it to ignore national and international rules like copyright. Wikipedia Foundation could be held responsible the larger community for flouting copyright, as music copying web portals were shut down by legal action after copyright violations. While Wikipedia seems an island, it connects to a social mainland that values ownership. Hence rather than force people to give edit rights away, Wikipedia could give item creators choices like:

1. *Public Edit:* Anyone can edit the item to improve it (default).
2. *Public Comment/Private Edit:* The item is open to comment by anyone, but only you can edit it.
3. *Private:* The item is viewable by others but only you can edit or comment.

Option 1 is currently Wikipedia’s only choice. Private rights (option 3) do not prevent a Wikipedia administrator from rejecting its display rights, i.e. “deleting” it.

Giving authors choice could open Wikipedia to many experts currently wary of it. In Wikipedia's criticism section Legio XX notes: "Dr MC Bishop, an archeologist and world renowned authority on Roman armor, wrote an article on the Roman lorica segmentata, only to see it mangled beyond recognition." and so refused to contribute. Andrew Orlikowski notes that well written articles are being "pecked" by amateurs until excessively long and frequently wrong. Wikipedia articles it seems can decay as well as grow. Choice lets a contributor topic expert give away some but not all control, and so not be overwhelmed by the majority, as the Wikipedia product is the last edit, and "revert wars" are won by the most persistent. Already within Wikipedia many suggest that "deserving" articles be "semi-protected", to limit allowed edits. "Private" contributions could be marked as such, to let readers evaluate credibility. Delegation gives even more complex options, as items could be open to the public for a while, then return to private editing. The general principle is that if Wikipedia wants to invite all authors, why not give authors freedom of choice, with public ownership as just one option?

Conclusions

Online societies like Wikipedia challenge humanity, asking what have we learned in several thousand years of society? If social knowledge can be put in information terms, computing could enable a global online society in the near future. If not, and if concepts like legitimacy have no computer meaning, then we must re-learn what social value means online. Wikipedia illustrates the struggle, as it began open and optimistic, then developed social structures and rules in response to vandalism. Its social rights model began simple but quickly became complex, and it could still fail as an online experiment by "social error". The difference between online and physical societies is that online "architecture" is defined by computer code, which in turn is defined by analysis and design. Rights analysis must become part of social-technical system design, to carry forward social knowledge into computer code, to close the "socio-technical gap" [1], and to help online communities succeed by increasing social health. Since the social level supercedes the technical one (Table 1), what a community says ought to happen actually should happen, not as an optional ethical "frill", but as a necessary requirement for social productivity. Wikipedia is new but its social problems are old, and communities over thousands of years have evolved social structures and rules as Wikipedia has done in just a few years. Social knowledge need not be relearned if we can define and discuss social rights in information terms.

We began with the premise that every information system object is owned, including the system itself. This seemed to make every online system a dictatorship, as indeed most bulletin boards are, albeit benevolent ones. However the social innovation of democracy suggests that in certain conditions, a group can "own itself" in general, i.e. the owner of a social-technical system can be its member community. Democracy, like privacy, can be seen as an extension of freedom, an individual's right to own him or herself. While the complexities of democratic voting cannot be discussed here, that online members of a system can own it brings our social logic full circle. Control "by the people" is fair, and "for the people" is socially beneficial, making democracy a legitimate solution to the social problem of who owns the

community. Democracy does not mean anarchy, as democratic community can still sanction individuals within it. Even Wikipedia, where anyone can edit anything, still is not entirely democratic, but one can see a social principle evolving, as after all, what will happen to it when its founder retires? The millennial IS challenge is to translate successful social rights into software code, because when online, code is law. The rights framework outlined in this paper can help meet that challenge, and provide a basis to teach a rights information analysis in social-technical system design classes.

References

1. Ackerman, M. S. (2000). The Intellectual Challenge of CSCW: the Gap Between Social Requirements and Technical Feasibility. *Human-Computer Interaction*, 15(2), 179-203.
2. Alter, S. (2001). Which life cycle --- Work system, information system, or software? *Communications of the AIS*, 7(17), 1-52.
3. Geen, R. G., & Gange, J. J. (1983). Social facilitation: Drive theory and beyond. In H. H. Blumberg, A. P. Hare, V. Kent & M. Davis (Eds.), *Small Groups and Social Interaction* (Vol. 1, pp. 141-153).
4. Grudin, J. (1990). The Computer Reaches Out: The historical continuity of user interface design. Paper presented at the Proceedings of CHI '90, ACM SIGCHI Conference, Seattle, Wash., USA.
5. Kuutti, K. (1996). Activity Theory as a Potential Framework for Human Computer Interaction Research. In B. A. Nardi (Ed.), *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge, Massachusetts: The MIT Press.
6. Lessig, L. (1999). *Code and other laws of cyberspace*. New York: Basic Books.
7. Locke, J. (1690). An essay concerning the true original extent and end of civil government: Second of 'Two Treatises on Government' (1690). In J. Somerville & R. E. Santoni (Eds.), *Social and Political Philosophy* (Vol. Chapter 5, section 27, pp. 169-204). New York: Anchor.
8. Rawls, J. (2001). *Justice as Fairness*. Cambridge, MA: Harvard University Press.
9. Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.
10. Whitworth, B., & de Moor, A. (2003). Legitimate by design: Towards trusted virtual community environments. *Behaviour & Information Technology*, 22 (1), 31-51.
11. Whitworth, B., 2006, Social-technical Systems, in *Encyclopedia of Human Computer Interaction*, Edited Claude Ghaoui, Idea Group Reference, Hershey. p533-541 <http://brianwhitworth.com/hci-sts.pdf>
12. Whitworth, B. and E. Whitworth, *Reducing spam by closing the social-technical gap*. *Computer*, 2004(October): p. 38-45.

Requirements Determination in a Community Informatics Project: An Activity Theory Approach

Melih Kirlidog

Computer Engineering Department, Marmara University, Goztepe 34722, Istanbul, Turkey
melihk@eng.marmara.edu.tr

Abstract. Requirements determination is arguably the most important phase of any system development project. This is due to the fact that the entire project is shaped according to the perceived or real requirements obtained in this phase. Although there is a wide body of literature on requirements engineering for IS development projects, there have been only a few attempts lately to theorize requirements determination for community informatics development. This article attempts to analyze the requirements determination efforts for a proposed community informatics project in Turkey by Activity Theory.

1 Introduction

Several system development methodologies have been developed for computerized Information Systems (IS). Avison and Fitzgerald define methodology as a collection of procedures, techniques, tools, and documentation aids which will help the developers in their efforts to implement a new system [2].

Although all methodologies have several steps, these steps can be classified in two major groups, namely analysis and design. A third group which may involve steps like testing and user training is the implementation phase where the final product is prepared to be put into service. This is a consecutive process, i.e. design cannot be performed before the analysis and implementation cannot be before the design. The name of a popular system development methodology, waterfall model [22] implies this consecutiveness. However, consecutiveness does not mean that there is no going back to the previous steps after a step is completed. On the contrary, system development is also an iterative process and it is a frequent requirement to return to the preceding phases and make modifications in IS development projects.

After an initial step that investigates the possibility and feasibility of a proposed project, the expectations from the system must be determined. This is accomplished by involving all or some of the stakeholders with the aim of determining what the system is supposed to do. Called requirements analysis, this is an indispensable stage in all formal and semi-formal system development methodologies. Since the requirements analysis step determines *what* is to be done in the succeeding steps of the development process, it can be regarded as the most important and consequential step in any IS design [8], [25]. As stated above, all IS development processes are iterative. This means that expectations about *what* the system must do may change

during the development process, implying that requirements determination continues during the lifetime of system development and beyond. Thus, not all, but the bulk of the requirements determination must be performed before the design.

As an emerging discipline, Community Informatics (CI) borrows several system development tools and concepts from traditional IS. This article attempts to analyze a step in requirements determination efforts in a proposed CI project in Turkey.

2 Activity Theory

Activity theory (AT) has its roots in German idealistic philosophy where Kant, Fichte, and Hegel studied mental activity (*tätigkeit*) in the process of establishing the relationship between subject (a person or a group of people) and object (motive of the activity of the subject). In the 1920s and 1930s a Russian psychologist, Lev Semyonovich Vygotsky developed the concept of tool mediation and object-orientation (nothing to do with object-oriented computer languages or databases) in psychology which was then dominated by psychoanalysis and behaviorism [26]. Although Vygotsky never used the term activity theory nor he did not make a thorough analysis of human activity, these two concepts became the most important tenets of AT which was later developed by his colleagues as a cultural-historical school of psychology. Leontiev introduced and further developed the term “activity” within the framework of AT [18]. Drawing on Vygotsky’s [26] “mediated act” which involves a stimulus-response process that also incorporates sign as auxiliary stimulus, Leontiev’s “activity” is not a reaction or set of reactions to a stimulus, rather it is a coherent and conscious endeavor with internal structure and transitions within that structure. This activity involves two interacting entities in the individual level, namely the *individual (subject)* and the *object*. The object is the aim or motive for the subject and their interaction is mediated by a *tool*. As a result of this interaction the *outcome* is produced after a transformation process. The motivation of the individual for performing the activity is the desire for forming the outcome. During the activity process it is possible that the subject, the object, and the tool can undergo some change. Tools mediate and are mediated by an activity. An object can be a tangible thing such as raw material for constructing a hut, or it can be an intangible thing such as an idea to construct a hut.

Although the two-level AT with subject, object, and tool has some explaining power in the individual level, it is inadequate for explaining complex activities that involve several individuals. Engeström extended AT to incorporate the community as the third level to address this inadequacy [9].

Beyond extending the model by incorporating community, Engeström’s model also encompasses mediating entities. Similar to the *tool* mediating *subject* and *object* in the two-level Leontiev model, *rules* mediate *subject* and *community*, and *division of labor* mediates *community* and *object* in Engeström’s model.

AT posits that an *activity* is meaningful with *actions*, and *actions* are meaningful with *operations*. Thus, we can talk about a hierarchy of human doings:

activities (motive) consist of → a chain of *actions* (goal) consist of → a chain of *operations* (conditions)

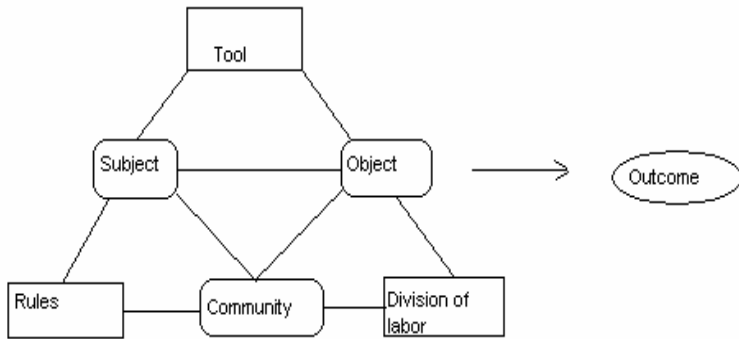


Fig. 1. Engeström's three-level structure of an activity

Activities, actions, and operations are not static structures; rather they can be interchangeable by the increased proficiency that comes by expertise. Leontiev gives the example of car driving for this process [18]. A novice driver's every operation is performed as an action with an *orientation* where *orientation* is the conscious planning phase of the action. In time, increasing expertise of the driver allows her to perform gear shifting unconsciously without any *orientation* and this crucial step in car driving becomes an *operation* for her in an action which might be increasing speed in a freeway or turning corner in a narrow street. Hence, shifting gears ceases to be an isolated goal-oriented process for the driver and it is performed unconsciously as if it did not exist.

Although individual operations and actions are themselves meaningful, they might seem irrational and meaningless in the same level when seen independently from a higher level. We can attribute a coherent meaning to the whole set only by investigating the question from a higher level. Leontiev's often-quoted example illustrates this fact [18]: In a hunt of a primitive tribe hunters are divided into two groups. While one group frightens animals by beating bush toward the catchers, the other group hunts them. Although seemingly irrational, bush-beaters' action is crucial to perform the hunting activity.

From an ICT point of view the most important tenet of AT is its tool mediation concept. Facilitating communication and information processing, computers are regarded as tools that mediate between subjects (human actors) and objects targeting an outcome. Having an important implication of social appropriation of ICT, this mediation is not merely between a person and the object, but also with other people. According to AT, tools can be tangible such as a pencil or they can be intangible such as a plan in mind. They are shaped by historical and cultural conditions and their usage shapes the way people behave. Hence, tools are carriers of culture and social experience. Tool mediation implies two interfaces. One is the interface between the user and the computer and the other is the interface separating the outside world from the user and the computer. Kaptelinin elaborates the concept of tool mediation by investigating the boundary between the individual with its tool and the external world [14]. He asks whether this boundary coincides with the boundary between the individual and the tool or with the one between the tool and the world. He states that

AT answers this question with the concept of *functional organs* which are “functionally integrated, goal-oriented configurations of internal and external resources”. External tools extend and complement human organs to accomplish tasks that cannot be accomplished by internal resources. For example, scissors elevate hands into an efficient cutting organ and notebooks enhance memory. Computers are also integrated into functional organs and they enable some enhancements in human capabilities such as practically unlimited memory capacity, efficient communication, and fast information processing. This applies not only in individual level, but also in the community level. In other words, computers can be functional organs that empower a community for, say, economic development, as much as they can be a functional organ for an individual fast and cheap communication.

3 AT and Information Systems

AT was introduced to the Western psychology in the 1960s and since then it became the most influential approach among contextualist theories of child development [24]. It diffused to the Human-Computer Interaction (HCI) sub-field of Information Systems (IS) discipline in the 1990s by the influential the work of Bødker [5]. During that time some Computer-Supported Cooperative Work (CSCW) researchers who were dissatisfied with the cognitive view of IT and HCI saw it too rational, too technology-oriented, and too individualistic. They were also appealed by AT which promised a social context. Thus, for those researchers focus of CSCW shifted from interaction of humans with computers towards focusing on the interaction of humans among themselves which is mediated by computers [1]. Beyond HCI and CSCW, several IS researchers who attempted to understand the nature of computer-mediated human activity in and out of organizations found AT useful for this task (e.g. [1], [12]). Hasan argues that AT offers real strength to IS research by providing cross-discipline discussion and by its suitability for descriptive and explanatory studies [11]. Hasan and Gould attempted to explain the ways in which senior managers make sense of, and use, organizational knowledge for computer-assisted decision making through AT [12]. They argue that AT is appropriate for studying “informaté-type” of systems (as opposed to “automate-type” [28]) like Decision Support Systems and Executive Information Systems, software systems used by middle or higher level managers. This is mainly due to the unstructured and messy nature of computer-assisted decision-making tasks of these people. Karlsson and Wistrand use AT to study behavior of actors in an IS development environment [15]. Berge and Fjuk use AT for understanding the roles of online meetings in a collaborated e-learning environment [4], and using AT, Scanlon and Issroff propose a new approach to the current practice in the evaluation of learning technology in the UK [21]. Fuentes et al. implement AT in studying the relationship patterns of Multi-Agent Systems in Computer Science [10]. They argue that these patterns guide analysis and design refinements and help to detect inconsistencies.

“The Connected Kids” project in Troy City of New York is an attempt to develop a CI project guided by AT [27], [13]. The project was conducted in Troy City which is a relatively poor area and it is host to the relatively affluent Rensselaer Polytechnic Institute where the conductors of the project are based. Thus, the background of the

project is a digitally divided environment. The action-research based project involves applying AT into the general design and requirements determination phase. The authors argue that when designing the project an important problem was to work with the potential users who had little computer literacy. AT had been a useful tool for solving this problem by distinguishing between the *actions* of individual users which do not seem meaningful without knowing what they are aimed and their *activities* which provide the context for understanding their objectives. Thus, AT enabled researchers to determine the real requirements even if they cannot be fully articulated by the potential users. The main method used in this process were the participatory design sessions that focused on the information necessary in users' work, family, and community activities and how this information might be provided by the system proposed.

At this point it is worth elaborating the difficulty of users in articulating their information requirements for a computer-based information system and how AT can provide a coherent guide to solve this problem. This is the most important difficulty in all IS development projects and, as stated above, it has its crucial implications for the later phases of the process as well as for the overall success of a project. Although there has been significant progress in software engineering, software development is mainly a labor-intensive craft [22]. This manifests itself particularly in the requirements determining phase where the most important task of the system designer is to "make the users talk" about their information requirements. Brooks, who was the project manager of the operating system development project for IBM System/360 computer laments that

"The hardest single part of building a software system is deciding precisely what to build. (...) the clients do not know what they want. They usually do not know what questions must be answered, and they almost never have thought of the problem in detail that must be specified." [6, p.199]

These remarks have been made in an organizational context where technicalities are strongly emphasized and users are highly proficient in ICT. Thus, they are even more relevant in CI projects where users are usually much less proficient in ICT. The explaining power of AT over human activities can be an ideal starting point for accessing the desired outcome of determining real user requirements in system development.

4 The Proposed Project

This study is based on an intended CI project in the town of Kars which is located on the Eastern border of Turkey. Town of Kars is the regional center of the province with the same name and the province has in excess of 300,000 inhabitants 90,000 of which live in the town. The main economic activity in the province is animal husbandry and agriculture. The town and its surroundings have a significant potential for tourism which is currently untapped.

Kars is one of the 81 administrative provinces in Turkey and it is one of the less developed ones in the country. According to the data provided by the Turkish State Planning Organization it occupies the 67. rank in the social and economic development index of Turkish provinces [23].

The project involved in this study is currently under investigation for TUBITAK (The Scientific and Technological Research Council of Turkey) support. It is a joint endeavor of Kars Municipality and the Istanbul Branch of Turkish Informatics Society (TIS). A unit of TUBITAK which is responsible for developing and supporting the Turkish distribution of Linux (Pardus) also supports the project. The project aims to contribute to the ICT usage in the town through education, providing free (or almost free) ICT access for the locals, developing some local ICT applications, and delivering some economic benefit to the region. The following activities are envisaged within the framework of the project:

- Educating children, women, youngsters, and teachers in the following topics with the education material that will be developed by the Pardus team:
- Developing a portal for the city that will contain the following:
 1. Developing a web site for fostering tourism that will have quality information about the cultural heritage and natural attractions in the town and surroundings.
 2. A web site that will have information about the events, bids, assets, and announcements of the Municipality. It will aim to provide total transparency in the Municipality activities. The Mayor and his assistants will be accessible through e-mail and/or chat.
 3. Establishing several e-mail discussion groups about the topics related to the town.

Table 1. ICT education program of the project

	Children	Women	Youngsters	Teachers
Estimated number of beneficiaries	1000	1000	1500	200
Computer supported education applications	√		√	√
Office applications (Introduction - Open Office)	√	√	√	√
Office applications (Advanced - Open Office)			√	√
Accessing knowledge sources on the Internet	√	√	√	√
Web design			√	√

- Providing some economic benefit to the town. Kars province is known for its high-quality dairy and other livestock products and it may be an option to market these products in the country through the Internet. Another alternative is to develop a sophisticated web site with an interactive and payment-enabled booking system for hotels in the province with the aim of fostering tourism. It is a crucial phase of requirements analysis to decide on one of these alternatives. Supporting an economic activity is seen as a measure to provide sustainability of the project after the funds cease.

If accepted, the project funding will last for 12 months. Although currently most of the members of the project team are from TIS, the project is designed in such a way that all phases will be carried out by the local sources of Kars Municipality and the

TIS team will act as coaches whose main task is to transfer technical knowledge to the local team while learning the local conditions from them. Transferring technical knowledge includes relatively sophisticated tasks such as interactive web programming. Hence, it is envisaged that the “foreign” contribution to the project will be as small as possible and ideally it should fade as soon as possible. This is seen as a remedy to avoid a “parachute project” [7] with little knowledge of local needs and goals. Such a project design is also important to overcome the positivist and technological determinant viewpoint widespread not only in the “field”, but also in TIS whose members are usually seasoned ICT professionals.

5 Requirements Analysis for the Project

Kars was visited twice by the project leader who is the author of this article. The first visit was in April 2006 and the second one was in early June 2006. Although the main objective of the first visit was to analyze the feasibility of the project, a rough requirements analysis was also performed which contributed to the document for funding application. The second visit was organized solely for determining the requirements. The most important task in this visit was to decide on the type of the economic activity to be supported by the project (selling dairy products on the Net or tourism). The activities in this second visit also shaped the main tenets of the project:

1. The project will be conducted as a Participatory Action Research (PAR) study. PAR aims to change a given condition to a desired state with the active contribution of all parties involved. Time and distance permitting and subject to the concerns about avoiding a “parachute project” mentioned above, members of the “foreign” and local project team will strive to engage as many locals as possible. This is true for all stages of the project such as requirements analysis, design, and actual implementation phases. “Serving two masters”, namely for developing a system that offers some benefit to the local population and attempting to understand the dynamics of a CI project in the local conditions, PAR is an ideal methodology for such projects [7], [3].
2. It is important to infer lessons for possible future projects. Although PAR is a robust research methodology that allows researcher to gain rich insight to the topic, it might be a better strategy to triangulate it with other methods such as survey instruments. Such a marriage of quantitative and qualitative methods yields even richer insight to the topic involved. To this end, two survey questionnaires, one in the beginning and one at the end of the funding phase of the project, will be distributed to the 604 employees of Kars municipality who can be regarded as a fair sampling of the town population. The questionnaires will contain almost the same questions that aim to gauge the quantity and quality of ICT usage in the town before and after the funded phase of the project. In other words, this longitudinal survey aims to investigate the benefits the project provided to the community during 12 months of its lifetime.

3. Unlike some other countries where CI projects are funded by the government, international agencies, and NGOs, there has been no such funding in Turkey. Consequently, there have been very few projects in the country which could be identified as CI [16]. Although private Internet Cafes are widespread and popular in the country (there are more than 20 in Kars town center alone where patrons usually have to wait in the queue to get a seat) there are almost no government-funded Internet access points open to public.¹ However, in the newly launched National ICT Strategy of Turkey it is envisaged to establish 4500 Public Internet Access Points during 2007 all over the country. Each having 20 computers and periphery devices such as printers and projectors for computer literacy training, these centers will be established in public buildings such as local libraries. The lessons learned in this project will be reported to the Directorate of Widespread Education in the Ministry of Education, the organization responsible for establishing and running these centers.

Requirements determination for such a project is a formidable task where several stakeholders are involved. Although it is ideal to engage as many stakeholders as possible in a PAR project, practical limitations such as time did not permit such a broad participation.

A total of nine interviews were conducted in the second visit. As noted above, that visit was organized solely for determining the requirements and the majority of interviewees were public officers. The two exceptions were the president of the local Chamber of Commerce and the manager of the Kars branch of Turkish Telecom which is a private monopoly since a few years. Although it sounds less than ideal to emphasize local bureaucracy and bypass other stakeholders in a PAR project, local conditions dictate such a measure. Turkey is a country with a strong bureaucratic tradition and it is important to get the support of the local bureaucracy and their concerns should be addressed in the project. Since it is obvious that remote control from Istanbul is not an ideal working environment for such a project and the project must be appropriated by the local owner, namely Kars Municipality, all interviews were conducted in the presence of Mayor's Assistant. He was also functional in arranging the interviews with the public officers.

Although the interviewees were generally positive about the project, some of them did not seem very enthusiastic. For example, the officer who is responsible for primary and high school education in the town had some concerns about the computer training in the project, because of the private computer education programs in the town which are being run by high school teachers. Such programs are useful and they provide some support to the modest salaries of teachers. It is quite probable that the project will adversely affect these programs. AT recognizes contradictions as

¹ A notable exception is the free Internet and computer access provided by the local libraries. This is facilitated by Ministry of Culture and it is quite widespread in the country. However, the benefit it provides to the local communities is doubtful. In the visit to the local library in Kars it was observed that all of the primary school children who were using the ten computers in the library were playing computer games and there was little doubt that the ones waiting in queue would do the same.

inevitable in human activity and one can expect more of them when the project actually starts. According to Kuutti contradictions in AT indicate a misfit within elements, between different activities, or between different developmental phases of a single activity [17]. In the above case the contradiction is the former, i.e. between teachers who teach in private computer courses and the project itself, and in order to avoid win-lose or lose-lose cases the project team must find creative ways to engage these people in the project constructively. Although not impossible, this is a difficult task in a community where social capital is low [20]. Indeed, low level of trust and self-organizing capability aspects of social capital are the main source of concern for the success of the Kars project where there are significant ethnic and political fault lines in the local community.

Unlike traditional system development methodologies, AT does not provide a detailed roadmap for requirements determination. Instead, it provides a useful insight to the requirement determination process as an object-oriented human activity with a clear target. The advantage AT offers is threefold. Firstly, it allows comprehending the requirements determination process as a logical and hierarchical progression. Related to this, it prevents getting lost in details. Thirdly, like the traditional system development methodologies, it allows determining outcomes –deliverables- quality of which can be used as a measure of success of the activity.

As stated above, the scope and nature of an activity is subject to change. An action can become an operation by increased proficiency that comes by expertise, or a time-dependent activity can be regarded as an action of a larger activity when looking at the big picture. Hence, although the task accomplished in the second visit can be regarded as an activity in isolation, it is in fact an action in the larger requirements determination activity. In a higher level of an abstraction, requirements determination activity can be explained by the AT in the following “Engeström triangle”:

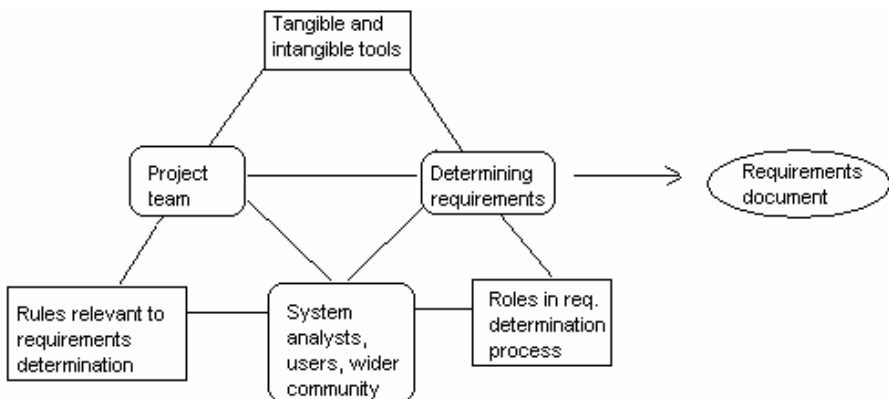


Fig. 2. Activity of requirements analysis

The elements in the triangle can have different meanings in different levels of activity. Analyzing the situation in two levels, namely the general requirements determination phase of the project and the requirements determination efforts in the

above mentioned second visit, the following table can be constructed. The table is by no means complete; for example, tools used in both activities can be extended to cover additional tangible utensils such as voice recorders in the interviews or intangible skills such as the interview organizing capacity of the Mayor's assistant.

Table 2. Hierarchical components of activities

	General requirements determination	Requirements determination effort in the second visit
Tool	Tangible: notebooks Intangible: Knowledge and skills of the analysts	Tangible: notebooks Intangible: Knowledge and skills of the analyst
Subject	All members of the project team responsible for requirements determination	Project leader and Mayor's assistant
Object	Determining requirements for the project that would provide maximum benefit to the local community	Determining the most appropriate economic activity that would be supported by the project
Rules	Local culture	Local bureaucratic culture
Community	Municipality, TIS, local bureaucracy, potential beneficiaries of the project	Municipality, TIS, local bureaucracy
Division of labor	TIS to organize and conduct interviews, Municipality to decide on the final form of the requirements	Project leader to conduct interviews, Mayor's assistant to organize them
Outcome	Requirements document	Choice for the economic activity

From the above table it is obvious that "requirements determination effort in the visit" column is a subset of "general requirements determination" column. In other words, hierarchical nature of tasks manifests itself as a nested sets of activities, actions, and operations in AT. This holds true for wider activities such as requirements determination as a subset of system development activity as well as individual steps within the requirements determination activity.

6 Conclusion

CI differs itself from traditional IS development activities in several aspects. Firstly, CI projects aim to provide some benefit directly for the user community whereas IS development in an organization usually aims to provide some benefit to the organization where the user community is only one of the stakeholders and *humans* are only one of the *resources* to realize organizational goals. Secondly, source-providing mechanism for system development in IS is the internal budget whereas CI projects are usually dependent on external funding. Thirdly, user population in CI projects for whom the project is developed tends to be less computer-literate than the professional user population in IS projects. And lastly, IS projects tend to be more focused and aim to solve a particular problem such as inventory control. On the other hand, a CI project can have several different and competing objectives such as computer training and providing some economic benefit to the user population.

These differences manifest themselves in the requirements determination efforts in both types. With a strong community orientation, requirements determination efforts

in CI development have to be different from the established requirements engineering of traditional IS development. With strong emphasis on the concepts of community, division of labor, and tool mediation, AT has a sound explaining power for requirements determination efforts of CI development. Armed with the theory, what is required at this point is to develop tools and methodologies for requirements determination and other CI development steps.

References

1. Aboulaflia, A.L.: The Cognitive and the Social Aspect of Computer-Mediated Work, Exemplified by the Research Traditions of HCI and CSCW. In: Hasan, H., Gould, E., Larkin, P., Vrazalic, L. (eds.) *Information Systems and Activity Theory: Volume 2 Theory and Practice*. University of Wollongong Press, Wollongong Australia (2001) 75-97
2. Avison, D.E., Fitzgerald, G.: *Information Systems Development: Methodologies, Techniques, and Tools*. Blackwell Scientific Publications, Oxford London (1988)
3. Baskerville, R., Myers, M.D.: Special Issue on Action Research in Information Systems: Making IS Research Relevant to Practice-Foreword. *MIS Quarterly* 28, 3 (2004) 329-335
4. Berge, O., Fjuk, A.: Understanding the Role of Online Meetings in a Net-Based Course. *Journal of Computer Assisted Learning*. 22, 1 (2006) 13-23
5. Bødker, S.: *Through the Interface-A Human Activity Approach to User Interface Design*. Lawrence Erlbaum, Hillsdale NJ (1990)
6. Brooks, F.P.: *The Mythical Man-Month: Essays on Software Engineering*. Anniversay Edn. Addison-Wesley, Boston (1995)
7. Day, P., Schuler, D.: Integrating practice, policy, and research. In: Day, P., Schuler, D. (eds.): *Community practice in the network society*. Routledge: London. (2004) 215-229
8. DeMarco, T.: *Structured Analysis and System Specification*. Prentice-Hall, Englewood Cliffs, NJ (1978)
9. Engeström, Y.: *Learning by Expanding*. Orienta-Konsultit, Helsinki (1987)
10. Fuentes, R., Gomez-Sans, J.J., Pavo, J.: Activity Theory for the Analysis and Design of Multi-Agent Systems. *Lecture Notes in Computer Science*. 2935 (2004) 110-122
11. Hasan, H.: Activity Theory: A Basis for Conceptual Study. In: Hasan, H., Gould, E., Hyland, P. (eds.): *Information Systems and Activity Theory: Tools in Context*. University of Wollongong Press, Wollongong Australia (1998) 19-38
12. Hasan, H., Gould, E.: Support for the Sense-Making Activity of Managers. *Decision Support Systems*, 31, 1 (2001) 71-86
13. Harrison, T.M., Zappen, J.P., Adali, S.: Building Community Information Systems: The Connected Kids Case. *Computer*. December (2005) 62-69
14. Kaptelinin, V.: Computer-Mediated Activity: Functional Organs in Social and Developmental Contexts. In: Nardi, B. (ed.): *Context and Consciousness*. The MIT Press, Cambridge, MA (1996) 45-68
15. Karlsson, F., Wistrand, K.: Combining Method Engineering with Activity Theory: Theoretical Grounding of the Method Component Concept. *European Journal of Information Systems*, 15, 1 (2006) 82-90
16. Kirlidog, M.: Developing Regional Communities in Turkey. In Marshall, S., Taylor, W., Xinghuo, Y. (eds.): *Developing Regional Communities with Information and Communication Technology*. Idea Group Reference, London (2006) 164-168

17. Kuutti, K.: Activity Theory as a Potential Framework for Human-Computer Interaction Research. In: Nardi, B. (ed.): Context and Consciousness. The MIT Press, Cambridge, MA (1996) 17-44
18. Leontiev, A.N.: Activity, Consciousness, and Personality. Prentice-Hall, Englewood Cliffs NJ (1978)
19. Nardi, B.: Activity Theory and Human-Computer Interaction. In: Nardi, B. (ed.): Context and Consciousness. The MIT Press, Cambridge, MA (1996) 7-16
20. Norris, P.: Democratic Phoenix. Cambridge University Press, Cambridge UK (2002)
21. Scanlon, E., Issroff, K.: Activity Theory and Higher Education: Evaluating Learning Technologies. *Journal of Computer Assisted Learning*. 21, 6 (2005) 430-439
22. Sommerville, I.: Software Engineering. 3rd edn. Addison-Wesley Publishing Company, New York Sydney (1989)
23. SPO: <http://ekutup.dpt.gov.tr/bolgesel/gosterge/2003-05.pdf> (2003)
24. Verenikina, I.: Cultural-Historical Psychology and Activity Theory in Everyday Practice. In: Hasan, H., Gould, E., Larkin, P., Vrazalic, L. (eds.) *Information Systems and Activity Theory: Volume 2 Theory and Practice*. University of Wollongong Press, Wollongong Australia (2001) 23-38
25. Vessey, I., Conger, S.: Requirements Specification: Learning Object, Process, and Data Methodologies. *Communications of the ACM*. 37,5 (1994) 102-113
26. Vygotsky, L.S.: *Mind in Society*. Harvard University Press, Cambridge MA (1978)
27. Zappen, J.P., Harrison, T.M.: Intention and Motive in Information System Design: Toward a Theory and Method for Assessing Users' Needs. *Lecture Notes in Computer Science*. 3081, (2005) 354-368
28. Zuboff, S.: *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books (1988)

Developing Enterprise Sponsored Virtual Communities: The Case of a SME's Knowledge Community

António Lucas Soares, Dora Simões, Manuel Silva, and Ricardo Madureira

INESC Porto, Rua Roberto Frias, Campus da FEUP, 4200 Porto, Portugal
asoares@inescporto, dsp@isca.ua.pt, mdasilva@iscap.ipp.pt,
ricardo.madureira@inescporto.pt

Abstract. This paper presents a case in the development of a knowledge community support system in the context of an industrial association group in the construction sector. This system is a result of the Know-Construct project which aims at providing association sponsored SME communities of the construction sector with a sophisticated information management platform and community building tools for knowledge sharing. The paper begins by characterizing the so-called construction industry knowledge community. The Know-Construct system concept and the its general architecture are described, focusing on the semantic resources, in particular the ontologies structure. The final part of the paper depicts the approach to the actual introduction of the system in the community. An action-research approach was planned to obtain research results regarding the social acceptance of semantic resources such as the ontologies and technical classifications used in system.

Keywords: enterprise sponsored virtual communities, ontology based information systems, action-research.

1 Introduction

Enterprise sponsored virtual communities (ESVC) are emerging as serious business schemes fostering collaboration and knowledge sharing both intra and inter-organizations. The community paradigm is winning space among more established inter-organizational interaction forms such as chains or networks, complementing them in some cases. ESVCs are complex socio-technical systems, difficult to design and maintain, needing multi-disciplinary approaches for their development. This paper¹ presents a case in the development of a knowledge community support (KCS) system in the context of an Industrial Association Group (IAG) in the construction sector. The system is a result of the Know-Construct project which aims at providing IAGs sponsored SME communities of the construction sector with a sophisticated information management platform and community building tools for knowledge

¹ Part of this paper result from research work financed by the Fundação para a Ciência e Tecnologia - MCTES - POCTI/GES/49202.

sharing. The Know-Construct project² intends to improve the effectiveness of the Construction Industry (CI) SME's by improving and extending the relationship with their customers through an innovative support regarding information and knowledge about products, processes and associated issues. This is achieved through specifically developed tools, supporting in particular the formation and operation of SME's knowledge communities in the context of Industry Association Groups (IAG). More specifically, these objectives aim (i) to provide a platform to support the creation and management of a community of CI SME's, coordinated by an association, fostering collaboration and knowledge sharing among its members (shareable knowledge includes, besides product and services information, companies' experience, e.g. best practices); and (ii) to provide problem-solving support to the individual IAG member's customers regarding the products' selection, applications and processes, as well as addressing other related problems such as legislative issues, safety issues, etc. This leads to a wider and deeper technical and professional competence shared by the SME's community, fundamental to its ability to satisfy customer needs, obtained through closer co-operation and knowledge exchange. It will be materialized as an internet-based platform that will offer the possibility to establish a "one to one" communications medium. Manufacturers and wholesalers (SME) may interact with their customers, advising them on specific topics relying also on knowledge created and maintained by the community of SME's mentioned previously.

These objectives were translated into two main modules of the KC system: (i) Customer Needs Management (CNM) System: a decision making support system regarding the products characteristics, applications and other consultancy services for SMEs customers applying a "web enabled dialogue"; and (ii) Knowledge Community Support (KCS) System: a system for SMEs to support a form of co-operation through the creation of Knowledge Communities of SMEs in Construction Industry.

2 Characterization of the CIK Community

There are many explanations around the community concept. After thorough review of the literature on this subject, a characterization of the Construction Industry Knowledge (CIK) community according to three approaches was made: type of utility, participant's behavior and typology.

The classification of different types of utility [1] presents a basic predictive model of different communities types, with particular relevance for those able to generate some type of utility for someone. The CIK Community can be classified as an hybrid of a Practice and Interest Community. On the one hand, company employees as individuals should see a direct utility to their particular jobs when participating in the CIK community. On the other hand this direct utility also comes into light when an employee (and consequently the company) realizes that, when solving a problem for an important customer, the information/knowledge used to reach the solution has been

² COLL-CT-2004-500276 KNOW-CONSTRUCT Internet Platform for Knowledge-based Customer Needs Management and Collaboration among SMEs (2005-2007). Project co-funded by the European Community. "Horizontal Research Activities Involving SMEs-Collective Research Programme.

made available by other community members. Nevertheless, not all the activities can be tracked to a causal benefit to the SME. For example, a chat session between two employees exchanging professional experiences or a report on a concern regarding the performance of a material in a news or blog entry by another employee, are activities that make sense in a community but cannot be assigned a concrete and immediate value for the organization.

Looking at professional development as the process of continually developing knowledge, skills and attitudes of professionals by means of formal and informal learning in the course of practice, the use of on-line knowledge communities for this purpose implies that an on-line knowledge community has to support this process. As a CIK community member, professionals in the construction sector will have a place for continual professional development that gives: individualized, flexible and easy access to a coherent and up to date knowledge domain, a range of opportunities to interact with like-minded persons and a range of opportunities to develop and exploit the knowledge domain. An example of this is: applying knowledge, learning from it, guiding others, disseminating ideas and results or doing research, embedded in a professional network. Our premise is that the membership of professionals of an on-line knowledge community will have positive effects on their continuing development, expressed not only in competences like knowledge, skills, experiences and attitude, but also, in the acquisition of organizational knowledge assets expressed in the growth and elaboration of professional knowledge, applicability of knowledge and legitimacy of knowledge. Based on work of [2], the CIK Community can be further characterized: (i) the goal is to develop and exploit knowledge about civil construction sector; (ii) there are continuous interactions between participants to meet these goals, (iii) information and communication processes are continuously made explicit, (iv) it adds value to the participants (professionals within the sector and customers alike), the on-line meeting place that is usable, (v) the culture focuses on the participants' needs as the route to high performance; involvement and participation create a sense of responsibility and ownership and, hence, greater commitment, and (vi) the context is highly complex and constantly evolving and the CIK Community will have to continuously cope with the expectations of its participants and their context of use of the system.

Based on the typology of virtual communities proposed by Porter [3] where the communities are classified under two levels - establishment and relationship orientation - the CIK Community is classified as an organization-sponsored community relatively to type of establishment and as a commercial community relatively to the relationship orientation. This community will have key stakeholders and/or beneficiaries (e.g. customers) that will play an important part in sponsoring the community's mission and goals. Being an organization-sponsored community, it will foster relationships both among members (e.g. professionals belonging to the associations of the project partnership) and between individual members (e.g. customers) and the sponsoring organizations (associations of the project partnership). Based on the classification of the CIK Community under the virtual community concept and the attributes commonly suggested in the literature to characterize virtual communities [4], the key attributes of the CIK Community. The key attributes that characterize of the CIK Community can be summarized as the Five Ps [3]: Purpose

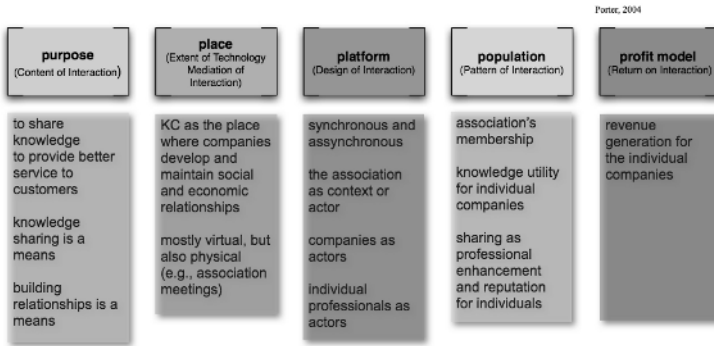


Fig. 1. Characterization of the CIK community according to the five Ps [3]

(Content of Interaction), Place (Extent of Technology Mediation of Interaction), Platform (Design of Interaction), Population (Pattern of Interaction) and Profit Model (Return on Interaction) (see Figure 1).

3 The KCS System Concept

3.1 The KCS System Functional Architecture

As mentioned before, the KC project has a very specific goal: to enable individual SME's to better solve their customers' problems. Therefore, KCS is focused on pursuing this goal in the first place. Although a knowledge community encompasses, as stated before, mechanisms that surpass this simple instrumental goal, the initial vision of the KCS system was specifically conceived having this in mind. This means that the KCS system supports CIK community building in a broad sense, though focused fundamentally in generating wide ranging and detailed knowledge to be used in managing the SME's customers' relationship, particularly in problem solving. The operationalization of KCS system is made through the use of mechanisms that will allow [5]: (i) the support of social processes (trust building, group formation and coordination), i.e., conditions for tacit knowledge exchange; (ii) increased levels of interactivity and to stimulate the dynamic exchange of knowledge (collaborative content management systems); and (iii) support to the personalization of user interaction (via the selection and presentation of content), maximize the impact of distributed knowledge and also facilitate the establishment of new relationships between the users. Keeping in mind the basic idea that the KCS system should support the CIK Community building in a broad sense, though focused primarily in generating a knowledge base that is as comprehensive and detailed as possible so as to be used in managing the SME's customers' relationship, particularly in problem solving, the following general functions of this module were specified.

Community building tools that support the processes of community building by providing the instruments to foster professional interaction and socialization; forums and weblogs are two such instruments and are tailored in KCS to be tightly integrated

with the semantic structure supporting knowledge management in KC; **Semantic resources management** i.e., the infrastructure and corresponding set of functionalities that support information and knowledge acquisition, organization and storage in KCS system, enabling (i) the management of classifications, thesauri and vocabulary, (ii) the acquisition of knowledge from digital content (including forums and weblogs entries, web pages, etc.) both internal to the CIK and from external sources, (iii) the maintenance of an ontology which is the base of knowledge representation, access and storage; **Knowledge resources access** i.e., creating, searching and updating knowledge resources constitutes a fundamental set of functionalities in KCS; although much of the community's information/knowledge will be created in communication/interaction processes (forums, weblogs), there will be also the need to create/access knowledge in a more structured way; digital content management and document management are the natural approaches regarding this issue.

This generic architecture can be decomposed in two layers (see Figure 2): KCS Core Services layer and Systems/Applications layer. KCS Core Services layer provides a set of services centred in the semantic resources management of KC. The basic architectural idea of KCS is to have a set of services to be used by specific, adaptable and, eventually, off-the-shelf systems/applications. The rationale is to take advantage of as great a number of open source systems/applications as possible that already provide the end user functionalities required in a knowledge community. For

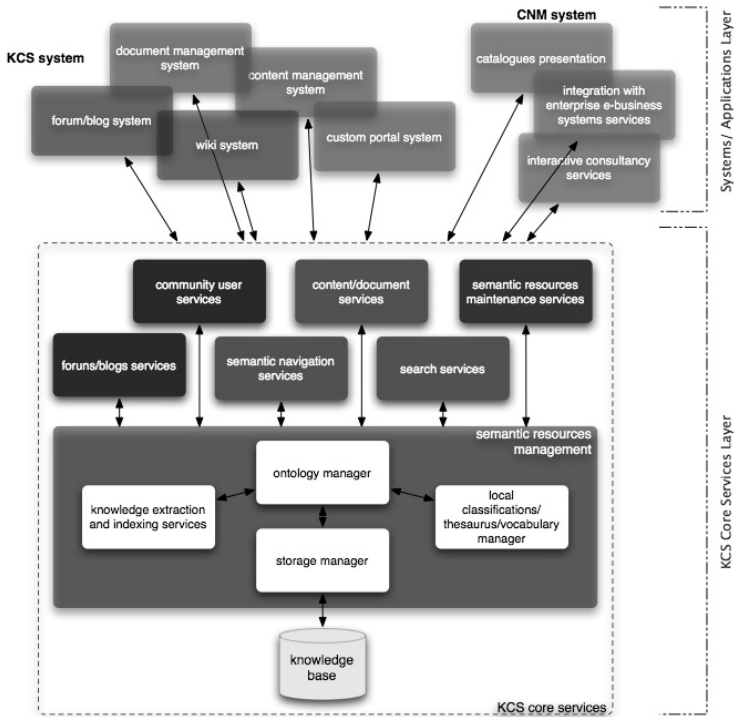


Fig. 2. Knowledge Community Support system

example, we can use a content management system (CMS) such as Zope/Plone or OpenCMS providing off-the-shelf functionalities to organize reports, data sheets, legal documents, and to publish web pages related with some community topic.

The CMS is configurable and extendable to use content/document services, search services and semantic navigation services in order to provide value added knowledge management to the community.

3.2 Ontologies Structure and Content

During the requirements analysis phase, a perception gained strength - to deal with the very concrete reality of the SMEs of each country and each Construction Industry Knowledge (CIK) Community of KCS, the KC system need to include a local ontology that would answer the KC partners' particular professional and cultural needs and attend to its social context of use. This meant the need to construct not one, but more than one ontology. Different countries (or even different associations) need to use local classifications, thesauri or ontologies. As different countries have different local ontologies, the best process to "plug-in" a local ontology into the CIK ontology is to align both ontologies. This has the advantage of making it easier to maintain and evolve the ontologies and, in practical terms, the option for ontologies integration can be a lot more difficult when managing the ontologies and its evolution process, for example when discovering the part of the integrated ontology that will suffer changes. The central ontology (CIK ontology) reflects standards and related classification schemes in the industry and the local ontologies will account for the individualized SME conceptual schemes, i.e. they will be strongly related to the consortium partners' needs.

The CIK Ontology that defines the domain of the CI summarized in the following sentence: The Construction Industry involves a set of *resources* that follow certain *conditions* which are used or required in a *process* that leads to *results*. As such, the proposed taxonomy includes four domains to classify these major concepts: Construction Resource; Construction Process; Construction Result; Technical Topic. The first three domains coincide with the major themes in the ISO 12006-2 standard [6]. The other domain (Technical Topic) is the result of the integration of an e-COGNOS module, further developed so as to include issues related to the CI that are not covered by the e-COGNOS ontology and IFC model. All domains integrate modules from e-Cognos ontology [7].

4 Implementing the KCS System in an IAG Sponsored Community

Empirical studies regarding the company's adoption and use of information systems based on advanced semantic infrastructures are scarce in the literature. Particularly when it comes to ESVC, our knowledge about the context, conditions and actual use of such type of systems is almost inexistent (as much as the authors are aware). In this project we decided to go further in the development of the Know-Construct system, taking advantage of prototype testing in realistic organizational situations to research the adoption and use of semantic enabled information systems in ESVCs. In this section we

will describe this approach emphasizing the research design and planning. This research is undergoing which means that existing results will be object of a future paper.

4.1 An Action Research Approach

Action research encompasses several different approaches. In this respect, Coghlan [8] distinguishes between mechanistic and organistic approaches. In addition, there are specific elaborations according to the type of inquiry, namely clinical, developmental, appreciative, cooperative, and participatory. These several ways of conducting action research may also reflect an emphasis on either the practical outcome [9] or the inquiry process [10]. In terms of the inquiry process, Avison [11] identifies problem diagnosis, change and reflection as the core activities of action research. Coghlan [12] similarly refers to planning, taking action, evaluating action and further planning. For the specific case of information systems research, Iversen [13] refers to initiating, iterating and closing, based on specific tasks that we synthesize below: 1. Establish the research team; 2. Identify the problem situation; 3. Delimitate the theoretical context; 4. Identify the research questions; 5. Plan specific tasks and coordination mechanisms; 6. Implement specific tasks and mechanisms; 7. Evaluate research and problem-solving; 8. Repeat cycle 5 to 7 as required to answer 2 and 4; 9. Exit cycle 5 to 7; 10. Elicit theoretical, empirical and methodological results. In practice, the action research's main dilemma is between researchers' control and participation. Sense [14], for instance, defines control as "the desire/actions of the researcher to influence the research proceedings in ways that 'drive' the process along the researcher's pre-determined path"; and participation as "the active engagement of the researcher in the activities of the study but without the need to dictate/mandate the processes in play". According to Sense [14] this dilemma can be attenuated by clarifying the roles of the participants and respective expectations while adjusting to changes in the project such as the mix of team members, the social, political and personal motivations, and the exact research questions and problem situation. In our view, this dilemma between control and participation can be further attenuated by ensuring a consensus on what constitutes the content (research questions and problem situation) and the context (theoretical and empirical) of the project. In other words, although the content and context of the study may change over time, the research team may still agree on the next task ahead. It is this permanent consensus that ultimately ensures the viability of action research.

4.2 Designing the Action-Research Based Approach

A first technical prototype of the KCS system was developed in two modules: (i) a minimally configured Content Management System providing basic content management, and basic community building functionalities i.e., a forum and a weblog application, and (ii) semantic functionalities module providing classification and ontology browsing, semantic searching, and content annotation. Two IAGs were selected to implement the first socio-technical KC prototype. Besides the IAG management unit, six companies in each IAG were chosen for the prototype testing and study. In each of these companies a set of people was selected to form the KC

Table 1. Specific topics to research through the KC prototype testing

semantic resources related	ESVC related
cognitive apprehension of the ontology and classification schemes	use of content shared within the community; trust and perceived importance of the information
intended and effective use of the ontology	intention to share information according a "community" worldview; actual uploading of content; quality of the shared information
contextual factors of ontology adoption: country, type of association, type of company	intention to use community building tools (forums, weblogs, wikis); actual use of community building tools

testing groups. During a training day they were introduced to the CIK community concept and to the use of the KCS system.

The general research questions can be formulated as the following: (i) how do workers in SMEs adopt and accept collaborative information management strategies, and (ii) how do workers belonging to SMEs participating in some form of network apprehend and appropriate the concept of enterprise sponsored virtual community. These research questions can be detailed in more specific topics (see Table 1).

4.3 Structure and Process of the Action-Research Intervention

Following the general guidelines from section 4.1 and the research questions formulated in section 4.2 the structure and process of the socio-technical prototype testing was designed, being depicted in the following figure:

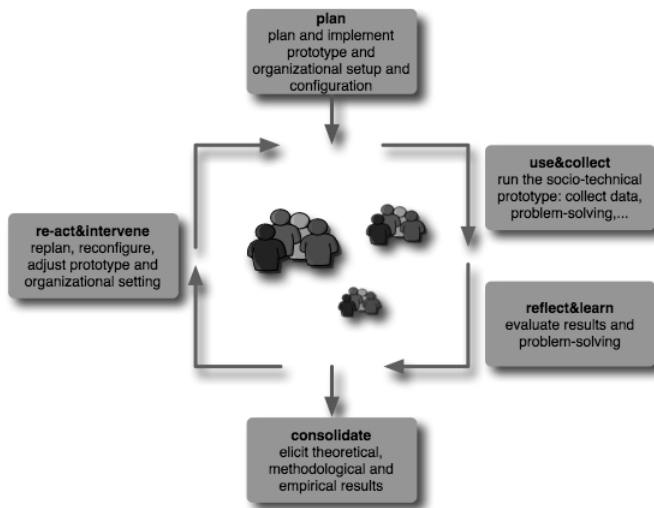


Fig. 3. The action research process of the Know-Construct prototype testing

5 Conclusions and Further Work

This paper presented the main issues in the development of a system to support a knowledge community of SME's in the Construction Industry. We presented the characterization of the Construct Industry Knowledge Community according to three approaches in the context of the Know-Construct project. Summarizing, the CIK Community was conceptualized as an aggregation of professionals and customers who interact around a specific shared interest of construction sector, sharing information and knowledge about products, services, techniques, legal aspects, experiences, etc, and where the interaction is totally supported and/or mediated by web technology and guided by some agreed protocols or norms.

To present the generic architecture of the KCS (Knowledge Community Support) system that supports the CIK community, the concept, layers and high-level design of KCS system were described. The high-level design organizes KC functionalities in four main areas: modelling (ontology development and use), tagging, query, and visualization. The future work will involve, obviously, the implementation and validation of the system. Further, one of the crucial development and validation tasks is related to the ontology implementation and the associated functionalities.

Given the selection, evaluation and structuring process described in this paper, it is highly probable that the CIK ontology reflects the standards and related classification schemes in the industry on the one hand and, on the other, that the local more specific ontologies will account for the individualized SME conceptual schemes, i.e. they will strongly relate to consortium partners' needs, as identified in the analysis of business case scenarios and in the users' requirement definitions.

The implemented method aims at developing a methodology of common Construction Industry Knowledge representation applicable to large sets of SMEs in the construction industry as a basis for the establishment of a knowledge community. Because of the available time frame, the described method was not as fine grained as desired. Therefore, further work in this area will be directed to detail the evaluation criteria. The next steps will involve the definition of the ontologies and maintenance strategies.

Although the phase regarding the analysis and specification of the KCS system (basic functionalities, semantic resources definition: high-level ontology and local ontologies definition, integration processes) has been comprehensive, and involved the users [15], we concluded that it was necessary to refine the requirements and design options through the use of a socio-technical prototype. This process is complex as we are dealing with virtual communities. Thus, there was the need of using innovative ways to establish a social test environment in order to achieve the goals of prototyping. We presented the design and planning for an action-research based prototype testing, which means not only to obtain new knowledge about the use of advanced semantic based tools for community support but also an innovation for the companies participating in the project.

References

1. Cornejo, M.: Utility, value and Knowledge Communities. Knowledge Communities. online:http://www.providersedge.com/docs/km_articles/Utility_Value_and_K-Communities.pdf , (2003). (last accessed January 2006)
2. De Vries, S. and P. Kommers: "Online knowledge communities: future trends and research issues." *International Journal of Web Based Communities* (1): (2004) 115 - 123.
3. Porter, C. E.: "A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research." *Journal of Computer-Mediated Communication (JCMC)* 10 (1), (2004) Article 3.
4. Blanchard, A.: Virtual behaviour settings: An application of behaviour setting theories to virtual communities. *Journal of Computer Mediated Communication*, 9(2) (2004).
5. Nabeth, T., Angehrn A. A.: *Towards Personalized, Socially Aware and Active Knowledge Management Systems. E-business and E-work - Challenges and Achievements in E-business and E-work*, Amsterdam, Holland, IOS Press (2002)
6. ISO 12006-2 Building construction — Organization of information about construction works - Framework for classification of information, DIS version 2001.
7. Lima, C., Fiès, B., Ferreira-da-Silva, C.: Setting up the Open Semantic Infrastructure for the Construction Sector in Europe – the FUNSIEC Project. In: 5th European Conference on Product and Process Modelling in the Building and Construction Industry – ECPPM 2004, Istanbul, Turkey (2004).
8. Coghlan, D.: Practitioner research for organizational knowledge: Mechanistic and organic-oriented approaches to insider action research. *Management Learning*, 34(4), (2003) 451-463.
9. Fricke, W. and Totterdill, P.: *Action Research in Workplace Innovation and Regional Development*. Amsterdam: John Benjamins (2004).
10. Reason, P.: Action research and the single case: A reply to Bjorn Gustavsen and Davydd Greenwood. *Concepts and Transformation*, 8(3), (2003) 281-294.
11. Avison, D., Lau, F., Myers, M. and Nielsen, P.: Action research. *Communications of the ACM*, 42(1), (1999) 94-97.
12. Coghlan, D.: Action research in the academy: Why and whither? *Irish Journal of Management*, (2006) 1-10.
13. Iversen, J., Mathiassen, L. and Nielsen, P.: Managing risk in software process improvement: an action research approach. *MIS Quarterly*, 28(3), (2004) 395-433.
14. Sense, A. Driving the bus from the rear passenger seat: Control dilemmas of participative action research. *International Journal of Social Research Methodology*, 9(1), (2006) 1-13.
15. Soares, A.L., Silva, M. Simões, D.: Selecting and structuring semantic resources to support SMEs, 8th International Conference on Enterprise Information Systems. Paphos, Cyprus. 2006.

Understanding Weblog Communities Through Digital Traces: A Framework, a Tool and an Example

Anjo Anjewierden¹ and Lilia Efimova²

¹ Human-Computer Studies Laboratory, University of Amsterdam, Kruislaan 419, 1098 VA Amsterdam, The Netherlands
anjo@science.uva.nl

² Telematica Instituut, P.O. Box 589, 7500 AN Enschede, The Netherlands
Lilia.Efimova@telin.nl

Abstract. Often research on online communities could be compared to archaeology [16]: researchers look at patterns in digital traces that members leave to characterise the community they belong to. Relatively easy access to these traces and a growing number of methods and tools to collect and analyse them make such analysis increasingly attractive. However, a researcher is faced with the difficult task of choosing which digital artefacts and which relations between them should be taken into account, and how the findings should be interpreted to say something meaningful about the community based on the traces of its members.

In this paper we present a framework that allows categorising digital traces of an online community along five dimensions (people, documents, terms, links and time) and then describe a tool that supports the analysis of community traces by combining several of them, illustrating the types of analysis possible using a dataset from a weblog community.

1 Introduction

Although research on online communities has a long-standing history, the technological infrastructure and social structures behind them evolve over time. In this respect communities supported by weblogs is a relatively recent phenomenon.

A weblog is “a frequently updated web-site consisting of dated entries arranged in reverse chronological order” [22]. Weblogs are often perceived as a form of individualistic expression, providing a “personal protected space” where a weblog author can communicate with others while retaining control [11]. On one hand, a randomly selected weblog shows limited interactivity and seldomly links to other weblogs [13]. On the other hand, there is growing evidence of social structures evolving around weblogs and their influence on norms and practices of blogging. This evidence ranges from voices of bloggers themselves speaking about the social effects of blogging, to studies on specific weblog communities with distinct cultures (e.g. [23]), to mathematical analysis of links between weblogs indicating that community formation in the blogosphere is not a random process, but an indication of shared interests binding bloggers together [17].

Often research on online communities could be compared to archaeology [16]: researchers look at patterns in digital traces that members leave to characterise the community they belong to. In the case of weblog communities relatively easy access to

these traces and a growing number of methods and tools to collect and analyse them make such analysis increasingly attractive (e.g. papers from the annual workshops on the Weblogging ecosystem at the WWW conference in 2004-06).

Many of the existing tools apply text or temporal analysis to large volumes of weblog data, often focusing on short bursts in time or popular topics (e.g. [1], [15], [21]). Others apply methods of social network analysis to identify and characterise networks between bloggers based on links between weblogs (e.g. [12], [19]). In our work we focus on combining both in order to go beyond currently available views on weblog data, aiming at developing tools that take into account existing community structures [14] and support the understanding of specific conversational clouds [18] and the “cloudmakers” behind them [20].

Although our work is based on the analysis of digital artefacts that weblog community members leave online, we find it important to articulate explicitly how studying the results points to more general questions about weblog communities: which digital artefacts and which relations between them are taken into account, and how the findings should be interpreted to say something meaningful about the community based on the traces of its members.

In this paper we present a framework that allows categorization of digital traces of an online community along five dimensions (people, documents, terms, links and time) and then describe a tool that supports the analysis of community traces by combining several of these dimensions, illustrating the types of analysis possible using a dataset from a weblog community.

2 Framework

In this section we present a simple framework that can assist in the analysis of online communities. The formulation of the framework is motivated by the perceived need to provide community researchers with a conceptual tool to focus on particular aspects of the community.

There is a strong relation between the framework we propose and ongoing research into the study of online communities. The field of social network analysis (SNA) can be characterised by studying the relations between persons and their links, sometimes taking into account time. The field of text mining from communities, sometimes called semantic social network analysis [4], mainly looks at the relation between terms and documents, largely disregarding the notion of the individual. Finally, the area of identifying trends in communities (e.g. [10], [8]) looks at documents, terms and time. The research that is closest to what we are trying to achieve is work on *iQuest* by Gloor and Zhao [9]. Their tool supports studying communities by making it possible to look at the community as a whole (topics discussed) and the contribution of members (who says what and when).

When thinking about online communities there are, therefore, at least five dimensions that play an important role and are possibly of interest for investigation:

Document. A self-contained publication by a member in the community. Examples of documents are a web page, email or weblog post.

Term. A meaningful term used by one or more members of the community. These terms occur in documents.

Person. A member of the community.

Link. A reference from one document to another document, and implicitly between the persons who authored the documents.

Time. The date, and possibly time, of publication of a document.

The framework thus focuses on communities that leave digital traces in the form of documents, and derives the other dimensions from the metadata (person, time) and content (terms, links). Given a dataset represented along these dimensions the researcher can navigate through it by specifying one or more initial dimensions, fixating a particular dimension (e.g. focusing on a particular term, person, or time period). Navigating along multiple dimensions makes it possible for the researcher to obtain both an overall view (what are the most frequent terms used in the community) and more detailed views (term usage of a particular member over time). The more dimensions involved, the more detailed, and maybe also the more interesting are the results of the analysis.

3 Tool

The framework has been implemented on top of a tool called tOKo [7]. tOKo is an open source tool for text analysis, with support for ontology development and, given the extensions described in this paper, exploring communities.

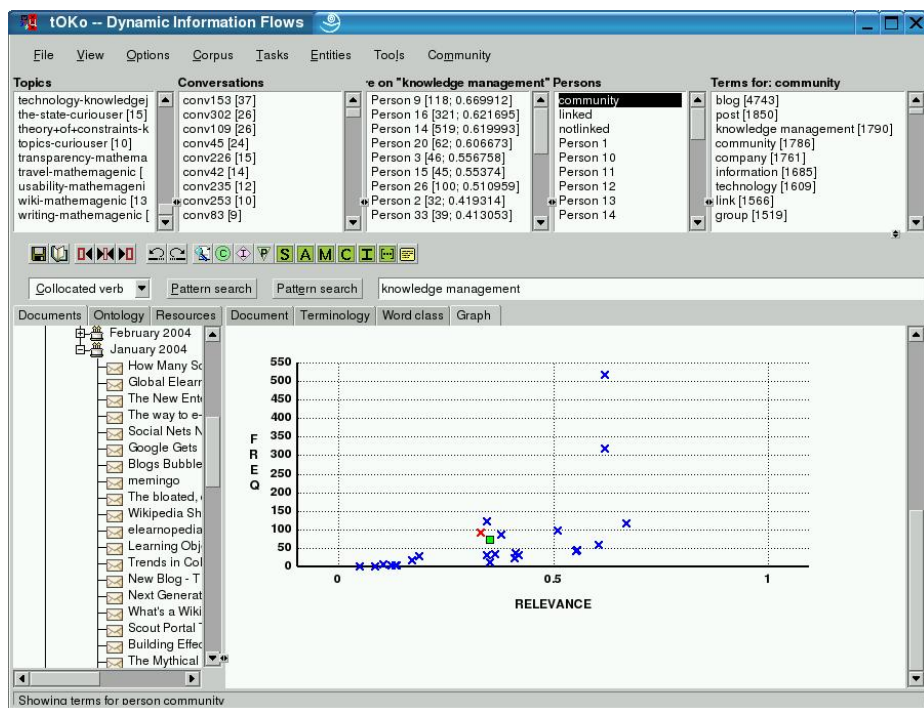


Fig. 1. Screenshot of the tOKo user interface with community research extensions

The only input tOKo requires is a corpus of (HTML) documents. Applied to community research, we assume a corpus represents the documents of the community under study. To be able to apply the framework it is necessary that for all documents the author (person) and the date of publication (time) is provided as metadata. Terms are automatically extracted by the *Sigmund* module of tOKo [2] which clusters lexical variants (abbreviations, inflections and alternate spellings) into a single, possibly compound, term. For example the term *community* has the lexical variants *community* and *communities*. Links between documents are by default extracted assuming the documents are HTML (anchor element). The implicit links between persons are inferred by considering the author of a document. After this extraction process is completed we have the data along all five dimensions for a given community represented by a corpus.

Fig. 1 shows the user interface of tOKo with the community research extensions.¹ The main difference between the base version of tOKo and the extension are additional methods and visualisations to cater for the links and time dimensions. The community research functions are available through the *Community* popup in the menubar and by clicking inside the five browsers with community data at the top.

4 Examples

Our research focuses on a cluster of weblogs in areas of knowledge management and social software. This is a dense social network of weblog authors, and may be classified as a community, given the many bonds and interactions between participants (see [6] for a discussion). The goal in this research is to understand how knowledge develops in a weblog community and to develop tools to monitor those processes. Although knowledge flows in a community are difficult to grasp with automatic analysis we focus on inferring them from combined investigation of patterns in language use and linking patterns between community members.

In the rest of this section we provide an example how the framework can be applied to address questions emerging in our study. In the analysis the tool is used to analyse a corpus of 6329 documents (weblog posts), over a period of a single year (2004), written by 24 persons (bloggers).

A researcher who wants to study a community asks herself questions. Answering such questions requires several transformations. First, a question will be related to a particular *slice* of the complete dataset according to the framework and the selection of the appropriate subset, which may result in less than five dimensions, is necessary. Second, on the resulting subset, some method of computation has to be applied. The particular method of computation obviously depends on the question asked, and in general the scientific community offers a wide variety of choices here. Finally, the results of the computation must be presented to the researcher. This is also the order we use in answering the questions below: relation to the framework, methods used and visualisation of the results.

4.1 Q: What Is the Topical Focus of the Community?

This is a question that can be answered by considering a single dimension: terms. The rightmost browser in Fig. 1 shows the most frequent terms for the community at hand,

¹ All other figures in this paper just show the content of the *Graph* sub-window.



Fig. 2. Network of terms that have a high co-occurrence with “instant messaging” in the community

low-content and high-frequency terms have been filtered out by *Sigmund*. A researcher might, based on this list conclude that this is a “knowledge management community”. The more frequent terms blog and post are the result of the publication medium.

A simple example of restricting the dataset is to only look at the terms of a single person. Some examples of the most frequent terms for individual members of the community are: **P1:** community, technology, virtual community, role, process; **P2:** knowledge management, organization, information, blog, article; **P3:** company, community, blog, corporate, knowledge management.

4.2 Q: How Are Community Topics Related to Each Other?

One approach to answering such a question is to consider an operationalisation of “related to” as “occurring in the same document”. This, first of all, requires two dimensions: document and term. And secondly, we need a statistical method to compute the relevance of what it means for two terms to be in the same document. For the latter we use the co-occurrence metric defined in [3]:

Definition: Let $n(B | A)$ (respectively $n(B | \neg A)$) be the number of occurrences of the term B in documents that contain the term A (respectively do not contain the term A), and likewise let $n(* | A)$ (respectively $n(* | \neg A)$) be the total number of terms in the documents that contain the term A (respectively do not contain the term A). Then the **co-occurrence degree** $c(B | A)$ is defined as

$$c(B | A) = \frac{n(B | A)/n(* | A)}{n(B | \neg A)/n(* | \neg A)}, \quad 0 \leq c(B | A) \leq \infty$$

We say that B co-occurs with A to at least degree k if $c(B | A) \geq k$. Note that $c(B | A) = 1$ if B is as frequent in documents containing as it is in documents not containing A , i.e. that term B and A seem to be unrelated.

Fig. 2 shows a graph of the co-occurrence network of the term “instant messaging” for the entire community. Edges between terms denote high co-occurrence (degree > 1.5). For example, “instant messaging” has a high co-occurrence with both “telephony” and “bandwidth”. In addition, “telephony” and “bandwidth” themselves also have a high co-occurrence with each other.

The co-occurrence metric thus provides a device that enables finding related terms in the community. Obviously, the same method can be applied to a member of the community by fixating a single point on the person dimension.

4.3 Q: Do Community Topics Change over Time?

This question involves the dimensions term and time. And the obvious way to answer it might be to plot the frequency of a term over time, similarly to what BlogPulse [8] does. There are, however, technical, social and perhaps principle reasons for not (only) using plain frequency over time. The technical reason is that frequency of term usage in a particular (small) community generally results in an irregular sequence of spikes that makes it difficult to identify trends. Trending frequency for very large communities (e.g. the entire blogosphere as with BlogPulse) does not have this problem: when a large event occurs, within days a significant proportion of the blogosphere will mention it. A social reason is that, because we are studying a community, there is a significant difference when a term is resonated in the community, compared to when it is not.

The social issue is addressed by considering that a community can be defined in terms of who-links-to-who (along the person dimension), but that the real discussions in the community in all likelihood consists of the linked documents. Put another way, there is a principle reason to view the community not just as the collection of all documents produced by the community, but to also look at the cross-section of the documents that are linked. In order to investigate this we split the dataset in two sub-sets depending on whether a link exists within the community as follows:

$$D_{linked} = d_i \in D \mid link(d_i, d_j), person(d_i) \neq person(d_j)$$

That is, D_{linked} is the set of all documents linked in the community excluding self-links. $D_{unlinked}$ is the difference between D and D_{linked} .

The technical issue is addressed by using *tf.idf* (term frequency vs. inverse document frequency) rather than frequency and by computing *tf.idf* over a sliding time-window to identify trends. The formula we use for determining the relevance of a term i for a document j is one of the many variants of *tf.idf*:

$$weight(i, j) = (1 + \log(tf_{i,j})) * \log(N/df_j)$$

where $tf_{i,j}$ is the frequency of term i in document j , df_j the number of documents that contain term i and N the total number of documents.

The time-window can be defined by the researcher, the default is a period of two weeks which takes into account that discussions in the blogosphere have a lag that is measured in terms of days rather than hours (as compared to discussions over email).

Fig. 3 shows an example of the trend of the term “knowledge management” used in the community. The x-axis represents time in days and the y-axis is the moving

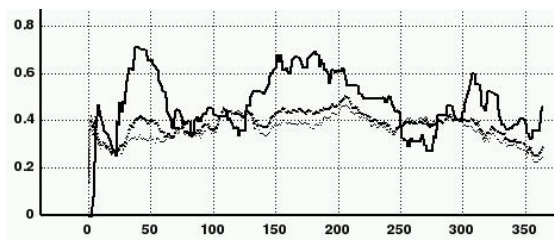


Fig. 3. Trend of “knowledge management” for all, linked and unlinked posts

tf.idf average over a period of two weeks. The three lines represent the values for the community as a whole (solid line), linked documents (D_{linked} , thick line) and unlinked posts ($D_{unlinked}$, dotted line). A conclusion that might be drawn from this visualisation is that “knowledge management” is more specific for linked posts than it is for unlinked posts and that this is the case continuously. Therefore, KM is one of the key terms of the community.

A graph for the term “Skype”, drawn using the same method, is shown in Fig. 4. The pattern which emerges is very different than the graph of KM. Skype is used in bursts from time-to-time (thick solid peaks), then dies away (unlinked usage is above linked usage). According to the terminology used by Gruhl et al. [10], in this community KM is a *chatter* term, whereas Skype occurs in *spikes*. Gruhl et al. look at all documents in their community and make no difference between whether links exist between documents. This corresponds to the average trend in the graphs and it is interesting to observe that both KM and Skype would be chatter topics using Gruhl’s approach, whereas the inclusion of the link dimension reveals that in our community the trend patterns are different between how KM and Skype are used in discussions.

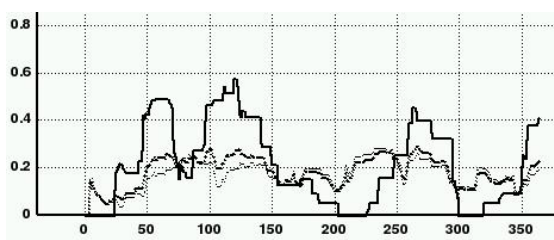


Fig. 4. Trend of “Skype” for all, linked and unlinked posts

4.4 Q: What Makes a Specific Community Member Different from Others?

We address this question by looking at term usage, other answers might come from studying linking practices (SNA). Intuitively, we can define that a person is aligned with the community when she uses the same terms at the same rate compared to the community as a whole. Terms that are used more often by a person (compared to the

community), might point to expertise or a certain passion for such terms (and the topics these terms point to).



Fig. 5. “Fingerprint” of term usage by a person with respect to the community

The method we use to find terms that significantly identify a person compared to the community is to compute for each term the following:

$$IDF(i) = \log(N/df_i)$$

$$sig(i, p_j) = average((1 + \log(tf_{i,j})) * IDF(i)), d_j \in D_{p_j}$$

where D_{p_j} is the set of documents by person p_j . In other words, we use the inverse document frequency (IDF) of the entire community and compare it to the average term frequency of a particular person. The results can be visualised by a list of the values for all terms. Fig. 5 shows a visualisation as a “fingerprint”, sometimes referred to as a *Zeitgeist*. In this kind of visualisation the font size of a term increases with significance.

4.5 Q: What Conversations Occur in the Community?

We define a conversation in a community as a set of documents that are linked directly or indirectly: if d_1 links to d_2 and d_3 links to d_2 all three documents are part of a single conversation [5]. Fig. 6 gives an example of the visualisation of such a conversation. Left to right is the time dimension, top to bottom are the persons involved in the conversation (in the tool the names are printed to the left of the boxes). The small coloured

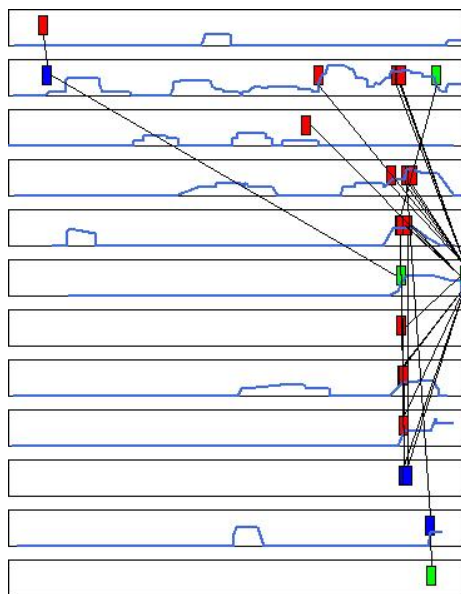


Fig. 6. An example of visualising a conversation

rectangles are the documents making up the conversation, and the lines between them are the links (it is only possible to link backward in time, so no directional arrow is required).

A related question that could be asked is: what is this conversation about? For this we use the same technique as with fingerprinting, and for this conversation, it results in the following terms being the most significant (compared to all documents in the community): KM Europe, personal knowledge management [workshop], keynote, workshop, Amsterdam, etc. The conversation was about a personal knowledge management workshop at KM Europe in Amsterdam. The researcher can guess this from the above list of terms or by inspecting the documents making up the conversation in more detail.

The wavelike lines for each person shows the use of the term “personal knowledge management” over time (*tf.idf* for the given person). As can be observed the second person mentions this term regularly, whereas some of the other participants only pick it up when entering the conversation. Such visualisations can perhaps be used to identify when the community picks up a new term and study the stickiness of terms over time.

Finally, we note that Fig. 6 contains datapoints from all five dimensions from the framework in a two-dimensional visualisation.

5 Conclusions

In this paper we presented a framework that allows categorising digital traces of an on-line community along five dimensions (people, documents, terms, links and time) and illustrated how it can be applied to translate general questions about a weblog com-

munity and its members into specific questions that could be answered with specific sub-sets of the data and specific methods for analysis.

In addition to supporting research on communities the approach we propose might be useful for community members, facilitators or sponsors. For example, newcomers are often overwhelmed by a wealth of information available in a community and find it difficult to navigate between multiple digital artefacts or find the right people to address. Dimensions of our framework could provide an additional way to navigate those, for example, by discovering documents or people associated with their own topic of interest. Community moderators or sponsors might be interested in identifying trends over time, to monitor patterns of activity (e.g. dynamics conversations) or to identify emergent “hot” topics or thought-leaders in order to adjust necessary support.

We find the framework and its implementation useful in several respects. First, it supports translating research questions into questions that could be answered based on a definite number of dimensions directly connected to the data available from the community interactions. Second, it inspires thinking of alternative ways to answer research questions by looking at combinations of the dimensions of data which might otherwise be missed. Third, it allows us to quickly compare alternate methods or research questions proposed in the literature. Finally, although our examples in this paper refer to a weblog community, our experience elsewhere suggests that the framework could be useful in other online community cases, for example for analysing forum-supported communities such as communities of practice.

Apart from extending the tool and its functionality we see two major challenges going forward. The main challenge is that we desperately need an “automatic” method to derive topics. Some researchers avoid this issue by simply stating term equals topic. Others are seriously addressing the topic issue, but report that automatically extracting them from weblog data is non-trivial [4]. Another challenge is to provide a user interface that allows a researcher to enter a question directly. Currently, the user interface is extended with a new control for each type of question.

Acknowledgements. The authors wish to thank Robert de Hoog and Rogier Brussee for their contributions to the underlying work and the reviewers for their constructive comments. This work was partly supported by the Metis project². Metis is an initiative of the Telematica Instituut, Basell and Océ. Other partners are CeTIM, Technical University Delft, University of Amsterdam, University of Tilburg and University of Twente (all in The Netherlands).

References

1. E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. HP Information Dynamics Lab, 2004.
2. A. Anjewierden, R. Brussee, and L. Efimova. Shared conceptualizations in weblogs. In Thomas N. Burg, editor, *BlogTalks 2.0: The European Conference on Weblogs (July 2004)*, pages 110–138, Vienna, February 2005. Danube University of Krems.

² <http://metis.telin.nl>

3. A. Anjewierden, R. de Hoog, R. Brussee, and L. Efimova. Detecting knowledge flows in weblogs. In Frithjof Dau, Marie-Laure Mugnier, and Gerd Stumme, editors, *Common Semantics for Sharing Knowledge: Contributions to 13th International Conference on Conceptual Structures (ICCS 2005)*, pages 1–12, Kassel, July 2005. Kassel University Press.
4. B. Berendt and R. Navigli. Finding your way through blogspace: using semantics for cross-domain blog analysis. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
5. L. Efimova and A. de Moor. Weblog conversations. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005)*, Los Alamitos, 2005. IEEE Press.
6. L. Efimova, S. Hendrick, and A. Anjewierden. Finding “the life between buildings”: An approach for defining a weblog community. In *Internet Research 6.0: Internet Generations (AOIR)*, Chicago, October 2005.
7. Anjo Anjewierden et al. tOKo and Sigmund: text analysis support for ontology development and social research. <http://www.toko-sigmund.org>, 2006.
8. N. Glance, M. Hurst, and T. Tomoyioko. Blogpulse. In *WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York, 2004.
9. P. A. Gloor and Yan Zhao. Analyzing actors and their discussion topics by semantic social network analysis. In *10th Conference on Information Visualization*, London, July 2006.
10. D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explorations*, 6(2):43–52, 2004.
11. M. Gumbrecht. Blogs as “protected space”. In *WWW Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, New York, 2004.
12. S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, and M. Tyworth. Conversations in the blogosphere: An analysis “from the bottom-up”. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS 2005)*, Los Alamitos, 2005. IEEE Press.
13. S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS 2004)*, Los Alamitos, 2004. IEEE Press.
14. M. Hodder. Link love lost or how social gestures within topic groups are more interesting than link counts. <http://napsterization.org/stories/archives/000513.html>, 2005.
15. M. Hurst. 24 hours in the blogosphere. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
16. Q. Jones. Virtual communities, virtual settlements, and cyber archaeology: a theoretical outline. *Journal of Computer-Mediated Communication*, 3(3), 1997.
17. R. Kumand, J. Novak, P. Raghaven, and A. Tomkins. Structure and evolution of blogspace. *CACM*, 47(12):35–39, 2004.
18. A. Levin. Conversation clouds. <http://alevin.com/weblog/archives/001692.html>, 2005.
19. C. Marlow. Investment and attention in the weblog community. In *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, 2006.
20. M. Ratcliff. Cloudmakers r us. http://www.ratcliffeblog.com/archives/2005/08/cloudmakers_r_u.html, 2005.
21. M. Thelwall. Blogs during the London attacks: Top information sources and topics. In *WWW Workshop on the Weblogging Ecosystem*, Edinburgh, 2006.
22. J. Walker. Weblog. *Routledge Encyclopedia of Narrative Theory*, 2005.
23. C. Wei. Formation of norms in a blog community. In S. Antonijevic, L. Johnson, C. Ratliff, and J. Reyman, editors, *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*. University of Minnesota, 2004.

Evaluating the Informational and Social Aspects of Participation in Online Communities

Emilie Marquois-Ogez and Cécile Bothorel

France Telecom R&D, 38-40, rue du Général Leclerc,
92794 Issy-les-Moulineaux, France
{Emilie.Marquois, Cecile.Bothorel}@orange-ft.com

Abstract. This paper aims at showing how beneficial it is to use an analysis grid as a guide to evaluate the dynamics of an online community and, on the basis of this grid, to develop agent-based models to explore each of the dimensions characterizing online communities within this grid. Several models from different perspectives have been proposed to study the dynamics of online communities (socio-technical perspective, life stages perspective...). In this paper, we consider two dimensions to evaluate them: the informational one and the social one. Their constituents are derived from an empirical analysis and a review of the literature on online communities and reputation.

Keywords: methodology, analysis grid, agent-based models, online communities, participation, informational aspects, social aspects.

1 Introduction

An online community can be seen as a complex system, the system being much more than the simple aggregation of its elements, which are “autonomous” entities with complex interactions. The study of any complex system and their composing parts need the use of a research method, especially if we aim at capturing and explaining complex behaviours. We propose in this paper to tackle a methodological point of view dedicated to the research field of social simulation in the particular case of its use to social system where social and informational aspects are of importance. Our method considers and organizes the different involved phenomena related to participation in online communities, proposes an analysis grid to express these phenomena, facilitates the choice of research directions. We also show how to use it in order to develop specific agent-based models to simulate the identified phenomena. Two main categories are considered in this grid to define participation in online communities. The first main category concerns the informational aspects (knowledge and interest) of participation in online communities whereas the second one concerns the social aspects (reputation, reciprocity and trust).

The paper is organized as follows. After presenting the different perspectives for studying the dynamics of online communities, we present our analysis grid, which is organized according to the two mentioned dimensions: the informational and the social one. We also give a synthetic view of the different concepts discussed (knowledge, interest, reputation, trust, reciprocity). In section 4, we describe briefly the collection of

agent-based models built using this grid. The last section explains how it is valuable 1) to use an analysis grid to understand how online communities work, and, 2) for each of their characteristics, to develop an agent-based model to explore it. This section describes too the encountered problems with this approach. The last section concludes presenting the main directions for future research.

2 Background

From the literature, several perspectives have been yet adopted to study the dynamics of online communities: sociological, socio-technical, life stages.

2.1 Detecting Communities in Networks

From a sociological perspective, the dynamics of online communities has been mostly examined from a structural point of view. Online communities are then envisaged or represented as social networks, i.e. as a set of people connected by a set of social relationships, the social network being represented as a graph.

The discovery and the analysis of communities in social networks have received a lot of attention by the mathematicians and the physicists in the recent years. The notion of community structure in networks was in particular pointed out by Girvan and Newman, defining it as “a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network.” [9]. Several methods for detecting the community structures in complex networks have been proposed and they all require “a definition of community that imposes the limit up to which a group should be considered as a community” [7]. The methods go from traditional methods such as spectral methods [11] [18] and hierarchical clustering [20], which are based on similarity measures, to more recent methods that include algorithms based, for instance, on betweenness measures [6]. The main differences between these methods concern the way community is defined and in particular the fact that the solution take into consideration multi-community memberships of an individual.

2.2 Sociability and Usability in Online Communities

From this perspective, online communities are socio-technical systems. They are supported by software technologies (Internet Relay Chat, newsgroup, bulletin board...) that mediate interactions.

In her book, Preece [19] gives some useful advices concerning community development and management. To improve the success of an online community, she proposes to consider two components, sociability and usability, in a single perspective; she argues that it is important to relate social interactions with interface design. The key components of sociability are the group's purposes and policies and the key components of usability are dialogs and social interactions support, information design, navigation and accessibility. Preece proposes a framework for usability and sociability that focuses on mapping the needs of people and their purposes to the policies and software that support them. Success depends on understanding the social and technical contexts in which the communication occurs.

2.3 Life Stages Perspective

From this perspective, the dynamics of online communities is considered through different steps, which represent each one of their steps of development. Three main steps are generally given: birth, maturity and death. The aim of these models is to understand how communities evolve in order to provide them with a better support and management.

The community's life cycle of Preece [19] is composed of four stages: pre-birth, early life, maturity and death. Pre-birth is when the development of the community is established, the software designed, and the initial policies are planned. During the second stage, new members are brought into the community and, although a diminution of the involvement of the developers, the community still needs nurturing. Maturity occurs when the community runs independently of direct guidance and when there is a critical mass of users. Death is the final stage of online communities, when members leave and the discussion slows down or ceases. McDermott [13] views communities of practice as living human institutions. He proposes a life-cycle perspective to describe the five stages (plan, start-up, grow, sustain/renew, and close) of the community development. This model is similar to Wenger's model but it associates the development of the communities with the understanding and the resolution of the tensions. According to Wenger [22], communities of practice pass through five stages (potential, coalescing, active, dispersed, and memorable), which are characterized by different levels of interactions and types of activities. The interactions generally increase through the active level and then decline through the dispersed stage, and pretty much disappears at the last one. Their model of Gongla and Rizutto [10] is similar to Wenger's and McDermott's in recognizing formative and growth stages of development (potential, building, engaged, active and adaptative). But it can't be considered as a life-cycle approach as a community can mature and dissolve at any one of these stages beyond the initial formation level. It is an evolution model, which describes how communities transform themselves, becoming more capable at each stage, while at the same time maintaining a distinct, coherent identity throughout.

2.4 Discussion

From the first perspective, researchers analyse email communication flow, visualize the structure and the dynamics of networks, delimit the frontiers of the communities, consider the number of links, try to detect communities in order to facilitate and improve collaboration or to help the understanding of the communication patterns within groups. The dimension considered to characterize online communities is structural. From the second perspective, the creation of meaningful interactions (i.e. the success of an online community) relies on the ability to communicate well. To communicate well, individuals need adapted tools. Online communities can be viewed through two distinct dimensions. The first one concerns the social aspects of online communities (purpose of the community, the needs of the individuals...). It is necessary to identify them. The second dimension concerns the technical aspects of online communities and consists in selecting tool(s) adapted to the social aspects, to facilitate the exchanges. The models included in the third perspective consider that an online community evolves and that at each step of this evolution, changes occur; for

instance, the number of individuals, the use of a new communication tool, the level of involvement of the moderator... They offer generic patterns of online communities. Each of the aspects is necessary to success in online communities.

According to us, these models have a limitation: except the model of Preece [22], which considers the needs of the individuals, these models consider online communities from a general point of view. The individuals who are apart of them are not really taken into account for explaining the dynamics. The models presented don't consider the factors that explain why an individual participates or not, don't focus on the motivations to contribute. Our proposal for explaining and understanding participation success in online communities, rely on personal attributes as we adopted and individual-based approach for explaining macroscopic behaviours. Thus, the dimensions proposed by the authors from the literature are useful but not sufficient.

3 Our Analysis Grid

Two sources of information have been used to elaborate our analysis grid: 1) the analysis of two-years (2002 and 2003) archives of the French-speaking mailing-list *ergo-ihm* (with 613 members, 698 e-mail address, 1880 discussion threads identified and 4101 exchanged messages), which is a community of practice in the area of human computer interface and human factors, 2) the analysis of the results of two surveys of individual interviews from the same mailing-list. The first one was conducted from July 11, 2004 to September 30, 2004 and answered by 88 members of the mailing-list. It deals with the activity of the participants to the mailing-list. A deeper questionnaire was conducted from February 16, 2005 to March 31, 2005 and answered by 23 identified members. It aimed to know if the interactions were motivated either by social or informational aspects. This sociological analysis allows us to identify different behaviours, which can be classified in two categories according to whether they can be explained by informational or social motivations. They constitute the two main sections of our analysis grid and are described in the next sections.

3.1 Informational Aspects

In this section, we distinguish the behaviours related to the information (asking advices for instance) we can find in mailing-lists from the motivations for reading messages and contributing, which are individual properties.

Information exchange typology. Different behaviours related to the information among online communities can be identified. The typology by Burnett [5] exhibits three kinds of behaviours, excluding lurking, hostile behaviours (trolling, flaming) and behaviours that are not specifically dedicated to information exchange (jokes, etc.); he distinguishes the announcements, the questions and other kind of information requests and group projects. The analysis of our surveys results and of the archives exhibits two main types of messages: post of messages that initiate a discussion (these messages can be divided in two primary types: messages that provide information such as call for conferences and job offers and requests such as questions and asks for references) and post of answers to requests.

Motivations for reading and posting messages. In an online community, interaction requires not only active posting, but also reading of others' messages. It is a pre-condition for sending answers. Members usually don't read all the messages. They make a selection according to the topic of the message. Are they interested in or not?

For members, posting messages is a way to help and inform the community of ergonomics (27 %). Participants answer a message for two main reasons: 1) they are interested in the subject of this message or 2) they have sufficient knowledge to share or they can improve or precise the previous answers. They make requests if they basically need of information. The members also want to be sure that their contribution will really be the answer the requester is waiting for. The behaviours associated to this informational dimension could then be seen as purely rational, members aiming at gathering useful information for them and answering requests in order to make the list survive for future needs.

3.2 Social Aspects

Reputation. We find in ergo-ihm the different categories of contributors identified by Millen and Dray [15]: regular contributors, sporadic contributors, very infrequent contributors and lurkers (individuals that never contribute or rarely). According to the participants in the survey, regular contributors are key members; the other participants consider them as experts. Moreover, their contributions are relevant and generally concern a specific domain. Reputation of the author of the messages influences the participation of other members. We consider here that reputation is a subjective measure and reflects how members perceived each other members. In our study, we consider several indicators to evaluate reputation of messages writers: the perceived knowledge, the perceived activity, the perceived citations, the reactivity, etc. From our inquiry, some members said that they were enabled to answer messages because if the perceived difference of knowledge between them and the author of the message was quite small, but they have hesitations when the author has a "high" reputation, i.e. appears to be an expert (considering here the level of knowledge as the main indicator).

Reciprocity. According to the analysis of the results of the second survey, the identity of the author of the messages influences the participation of the readers. For example, some members said that they answered more readily to members that helped them previously or helped others members. Moreover, to the question "which type of member do you appreciate?", they essentially say that they have a preference for the members who contribute and particularly the members that participate in a relevant way.

Trust. "At any given time, the stability of a community depends on the right balance of trust and distrust. Furthermore, we face information overload, increased uncertainty and risk taking as a prominent feature of modern living. As members of society, we cope with these complexities and uncertainties by relying trust, which is the basis of all social interactions" [1]. 80,7 % of the respondents considers that trust is a necessary condition for a community to live. Trust helps a community to better work. Focusing on the messages, the exchanged information can be viewed in terms of: richness, exactitude, integrity, relevance, interest, mutualisation, quality, credibility, honesty, guaranty, respect,

conviviality, truth and legitimacy. If that trust doesn't exist, disappears or is abused, the community suffers (there are less posts, the members unsubscribe), and may indeed collapse. Some respondents precise that all the members can't be trusted in the same manner. According to some members, trust is not important: the information that is exchanged in the list is general, is only recommendations. Moreover, it is to anyone to take what he wants to take, knowing that each of the informations can be verified. Like reputation, trust is a very subjective perception about each other and also about the global list.

3.3 Our Analysis Grid

On the basis of this empirical analysis, we developed an analysis grid compiling the described criteria above. We will use it as a guide to explore participation in online communities. Our grid is composed of two main categories that cover the various aspects of the informational and social dimensions of participation in online communities. We don't consider in the grid emotional factors although they do influence the dynamics of the mailing-list and the nature of the relationships. For instance, some of the respondents said that the tone used in the messages had an influence on their participation. Others answered that they couldn't help some members because these ones had had a bad behaviour with some members (for instance, new members or students). We can also precise that social events are sometimes organized which can modify the exchanges between members. Emotional factors are numerous and various and it seemed to us that it was first more relevant to consider only two dimensions in order to not complex more our model. Moreover, they are difficult to model. Fig. 1 shows a schematic representation of the grid.

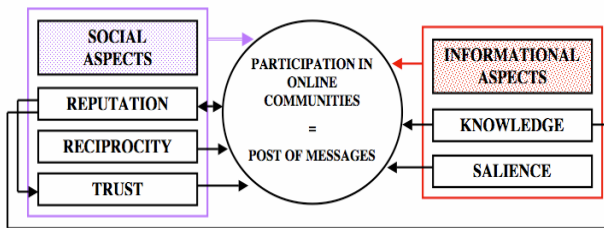


Fig. 1. Schematic representation of the analysis grid

Participation in online communities consists in posting messages and can be viewed as a form of information sharing. Informational aspects and social aspects influence participation. Informational aspects can be decomposed into the knowledge an individual has on the subjects and the saliency he accords to each one. Social aspects that influence participation are 1) the reputation an individual gained from his level of participation and his level of knowledge, 2) the reciprocity (we refer to the interchange between two individuals and to the reciprocating toward individuals indirectly through a third party), and 3) trust, which corresponds to the probability that an individual will perform a specific action. Notice that reputation influences trust.

4 From the Analysis Grid to Agent-Based Models Development

As a complementary tool to the analysis grid and the classical sociological analysis, we use agent-based social simulation (ABSS). It enables to describe observed phenomena but also to propose formalized hypothesis concerning the generating mechanisms [14] of these phenomena (by the simulation) at an individual level (following the multi-agents approach) or of regularities observed at a macro level [3][8].

4.1 Scenarios Elaboration

On the basis of the grid, we defined several scenarios in order to explore the different aspects observed in the real data. Some of them are presented in Table 2. We focused first on the informational dimension of online communities by combining the knowledge an individual has on a subject and the interest he gives to it.

Table 2. Description of the rules applied for three scenarios – concerning the calculation of probaInit, probaAns and levelKnewMess / probaAns: probability to answer a message; probaInit: probability to initiate a discussion; probaRead: probability to read a message; levelKnewMess: level of knowledge contained in a new message

	Probabilities			levelKnewMess
	probaAns	probaInit	probaRead	
Scenario A	static [0.01,0.1] – 0.01	= levelS probaInit > nb	= levelS	-
Scenario B	= (levelKmess – levelKMemb)	= levelS probaInit > nb	= levels	[0,Kmemb]
Scenario C	= (levelKmemb – levelKmess)	= levelS probaInit > nb	= levelS	[0,Kmemb] [Kmess,Kmemb]
...

The changes from one scenario to the other concern the calculation of the probability to answer, the calculation of the probability to initiate a discussion and the calculation of the level of knowledge contained in the messages. For instance, in the first scenario (scenario A), described in Table 2, the probability to answer a message about a subject corresponds to a static value, which is the same for all the individuals, whereas the probabilities to initiate a discussion on a subject and to read a message about a specific subject correspond to the salience an individual gives to the subject. In the first scenario, we don't consider the level of knowledge an individual has on subjects nor the level of knowledge contained in the new messages.

The next step will consist in creating scenarios for exploring social aspects of participation in online communities and in combing different scenarios that take into account both informational and social aspects. In the next section, we propose an agent-based implementation to test our different scenarios. We describe the models we implemented.

4.2 Experimentation Models

In this section, we present briefly the different agent-based models we developed, following the structure of the analysis grid (Fig. 2). For each model, we formulated hypothesis and various declinations on each type of model.

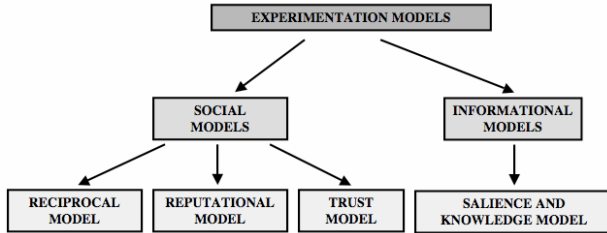


Fig. 2. Hierarchy of models

Information model. Interactions depend on the knowledge a participant has on a subject and/or the salience he accords to it. He will post an initiating message on a subject if he has information to give and will send a request about a subject if he needs some. He will answer a message if he has enough knowledge or if he can bring something new to the previous answer. Notice that for some of these informational behaviours, the agent doesn't need to be interested in the subject, and when he asks for some help concerning a subject, the subject generally interests him.

Social models. We distinguish three social models: the reputation model, the reciprocal model and the trust model. For each model, we formulated hypothesis.

Reputation model. The reputation of an agent concerning a subject depends on his knowledge and on his level of participation (how many messages did he post?) on it. Notice that there is as much evaluations of individuals than of members in the mailing-list. In this model, three types of participating behaviours relying on reputation are considered: an agent contributes only to increase his reputation; an agent read a message about a subject according to the reputation of its author; an agent answers essentially to messages whose author has a high reputation.

Reciprocal model. In this model, the participation of an agent depends on the participation of the others. An agent prefers to answer a message sent by an agent who already helped him (direct reciprocity) or helped the others (indirect reciprocity). Direct reciprocity refers to interchange between two agents and indirect reciprocity "refers to interchange between two concerned agents interceded by mediating agents in between." [19]

Trust model. In this model, an agent will place trust in another agent according to his reputation. Trust not only concerns the relations between the agents but also the trust an agent places in the mailing-list.

5 Discussion

Online communities can be viewed as complex systems: this means that to understand how they function, it is necessary to use a methodological approach, both precise and rigorous, which allows exploring the multi-facets of online communities' participation, by making adapted and relevant experimentations.

In this work, we used an analysis grid and a related collection of models to study participation in online communities. Why did we use this tool and this approach? That's what we explain in this section.

Elaboration of an analysis grid. Several authors [4] [12] [21] defined modeling processes. These researchers agreed that it is necessary to first formulate the problem (it consists in defining the goals of the study and in determining what needs to be solved), next, to plan the project in detail (the objective of this step is to ensure that sufficient time and resources are available for completion), and then to define the conceptual model. Law and Kelton [12] and Banks and al. [4] explain that data collection is an essential step in the model conceptualization, but don't explain how we must use these data, describe them, organize them in order to define the conceptual model? We consider that these elements are important. How to organize the information extracted from the data? Which aspects should we consider? How many models do we need to implement to test hypothesis? We proposed, in this work, to use an analysis grid.

Our analysis grid aims at answering to those questions. It is linked to problem formulation and consists in the result of the data analysis. It is a kind of guide useful for leading experimentations. In this grid, the main dimensions that characterize the systems and the various aspects that are used in these categories are identified by a name. Each category is also defined and illustrated by examples.

By using this tool, we are able to evaluate the need to develop several models, and then to help us to develop them. An analysis grid has several advantages: it joins together in a single tool the main information to consider the models to develop (they are identified by a name, a definition and a set of examples), organizes and structures the main components of the system, presents in a relevant way the main mechanisms to consider in the model(s). This approach provides a mean to discover the relevant aspects and relate them in a relevant way.

Collection of models as a framework. On the basis of the analysis grid, we developed a collection of models, which account for the complexity of our social system, the mailing-list ergo-ihm. We have chosen this approach because we think that several models are richer in information than only one: a single model allows to answer only a few questions whereas a collection of models, by maintaining a multi-levels view of the system, allows to observe how the system functions at different levels of abstraction. Other advantages can be associated with this approach: it allows to exclude complex models and thus to avoid too much details and "enables to validate more securely the final model, because it furnishes a more complete depiction of the real system, at different grains" [2]. One can precise too that we used an agent-based modelling approach. Agent-based modelling is a powerful approach "because it combines the rigor of formal logic with the descriptiveness of an agent

paradigm for representing social actors and their interactions” [18]. It can bring valuable insights thanks to its flexibility in the construction of agent architectures.

6 Conclusion

A participation analysis grid has been elaborated on the basis of a sociological analysis to help understanding participation in online communities. It is composed of different categories and gives a view of the different factors that may influence participation. Associated with agent-based modelling, this tool gives valuable insights.

On the basis of this grid, based on an empirical analysis concerning participation within mailing-lists, we built a collection of agent-based simulation scenarios. Our objective was to test hypothesis concerning the generative mechanisms of the behaviours isolated from data analysis. A first probabilistic scenario (scenario A) dealing with the way participants answer to messages highlighted a burst in messages due to the following rule: the more responses there are, the more messages to answer to. The analysis of the results let us suppose that the change of the static variable by different equations taking into account the level of knowledge contained in the message and the level of knowledge of the agents could allow controlling this burst. This hypothesis was verified in the two following scenarios (scenarios B and C). But, we have to remark that even the hypothesis we took for scenarios B and C, can be realistic in some cases, they cannot be applied systematically, and surely the case where population is heterogeneously composed of agents having one or the other behaviour has to be studied.

Another step consists in considering the reputation, reciprocity and trust effects in the behaviour of the individuals on mailing-lists in order to concentrate on the social dimension of the phenomenon. We also plan to consider several subjects (the levels and the types of participation can vary from one subject to another one for a unique agent).

References

1. Abdul-Rahman A., Hailes S.: Supporting Trust in Virtual Communities. Hawaii Int. Conference on System Sciences, Vol. 33, Maui, Hawaii, (January 2000).
2. Amblard, F., Ferrand, N. et Hill, D.R.C.: How a conceptual framework can help to design models following decreasing abstraction. Proceedings of SCS 13th European Simulation Symposium, Marseille, (octobre 2001) 843-847.
3. Amblard, F., Phan, D.: Modélisation et simulation multi-agents pour les Sciences de l'Homme et de la Société : une introduction, Hermès, to appear,(2006).
4. Banks, J., Carson, J. S.: Discrete-Event Simulation. Prentice-Hall International Series, London, (2001).
5. Burnett, G.: Information exchange in virtual communities: a typology. Information Research, Vol. 5, n°4 (2000).
6. Castellano, C., Ceconi, F., Loreto, V., Parisi, D., Radicchi, F.: Self-contained algorithms to detect communities in networks. Eur. Phys. J. B, Vol. 38, n°311 (2004).
7. Duch, J., Arenas, A.: Community detection in complex networks using extremal optimization. Phys. Rev. E, Vol.72 (2005).
8. Gilbert, N., Troitzsch, K.: *Simulation for the social scientists*, Open University Press (1995).

9. Girvan, M., Newman, M. E. J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, (2002) 7821–7826.
10. Gongla, P., Rizzuto C. R.: Evolving communities of Practice, IBM Global Services Experience. *IBM Systems Journal*, Vol. 40, n°4 (2001).
11. Kernighan, B. W., Lin, S.: An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, Vol. 49, (1970) 291–307.
12. Law, A. M., Kelton, W.D.: *Simulation Modeling and Analysis*. McGraw-Hill, Inc., (1991).
13. McDermott, R.: Community Development as a Natural Step. *Knowledge Management Review*, Nov-Dec. (2000).
14. Manzo, G.: Variables, mécanismes et simulations. Une combinaison des trois méthodes est-elle possible ? Une analyse critique de la littérature?. *Revue Française de Sociologie*, vol.46, n°1 (2005).
15. Millen, D. R., Dray, S. M.: Information Sharing in an Online Community of Journalists. *Aslib Proceedings*, Vol. 52, n°5, (2000) 166-173.
16. Moss, S.: Applications Centred Multi Agent Systems Design (With Special Reference to Markets and Rational Agency). International Conference on Multi Agent Systems (ICMAS-2000), Boston MA, IEEE Computer Society (2000).
17. Mui, L., Mohtashemi, M., Halberstadt, A.: A Computational Model of Trust and Reputation. 35th Hawaii International Conference on System Science (2002).
18. Pothén, A., Simon, H., Liou, K.-P.: Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, Vol. 11, (1990) 430-452.
19. Preece, J.: *Online Communities: Designing Usability, Supporting Sociability*. Chichester, UK: John Wiley & Sons (2000).
20. Scott, J.: *Social Network Analysis: A Handbook*. 2nd ed., Sage, London (2000).
21. Shannon, R. E.: Introduction to the Art and Science of Simulation. Proceedings of the 1998 Winter Simulation Conference. Association for Computing Machinery, New-York, (1998) 7-14.
22. Wenger, E. *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press (1998).

An Approach to the Assessment of Applied Information Systems with Particular Application to Community Based Systems

Driss Kettani¹, Michael Gurstein², Bernard Moulin³, and Asmae El Mahdi¹

¹ School of Science and Engineering, Al Akhwayan University, Ifrane, Morocco

² Center for Community Informatics Research, Development and Training,
Vancouver, Canada

³ Computer Science and Software engineering Department, Laval University,
Quebec, Canada

Abstract. This paper presents an empirically based method of indicator measurement for systematically linking applied Information Systems project outcomes with their technology inputs and particularly for assessing the situation before and after the development and deployment of the system. The result is a method to guide technology developers and decision makers in developing and deploying systems for which the community based outcomes may be systematically assessed and specifically in this instance as applied in the development and deployment of e-Government systems in support of enhanced locally based e-good governance.

Keywords: applied information systems project outcomes, systematic assessment, systematic measurement, community based outcomes, indicators, e-Government systems, good governance.

1 Introduction

A major driver underlying the process of globalization is the expansion of communication technologies: “computer networks, telephony, electronic mass media, and the like” [1]. The expansion of these technologies has changed the face of the world by enabling immediate interactions across political borders and irrespective of geographical distances [1]. Such expansion enabled the penetration of these technologies into many institutional structures, including as in the current case, public administrations in the form of “E-Government.” In this context there is an increasing interest worldwide in implementing E-Government projects recognizing that ICTs have been found to transform states in ways that positively influence their quality of governance as is the underlying objective of the ongoing Fez E-Government Project.

As background to an action research project supportive of local e-Government it has been necessary to review the current range of understandings with respect to certain of the concepts of particular significance. These notions include those of governance, e-governance, and “good” governance. Our intention through this analysis is to identify means to operationalize certain of these concepts as a way to systematically assess the

significance and value of the outcomes being produced. This paper presents the outline of this methodology for empirically assessing the outcomes of e-Government implementations through the measurement and evaluation of the outputs of these implementations and in the context of their contribution to broader governance objectives. This analysis is done in the context of assessing the outcomes of an E-Government project currently underway in Fez, Morocco specifically as a means to measure the contribution of the project to enhanced good governance by examining the governance situation before and after the development and deployment of the system.

In this process a method has been developed for linking the broad “vision” of a project’s proponents with project outcomes and related indicators. The overall intention is to provide a well-defined method of analysis as a means to guide developers and decision makers that want to develop and deploy e-government (or other systems with socially purposeful goals) in pursuit of identified governance (or other) enhancements. This paper presents that method.

It will of course, be noted that the specific content of the methodology being employed varies directly with the objectives of the system being implemented and even more specifically in relation to the goals of the particular stakeholders on behalf of whom the assessment is being undertaken. That we have chosen to highlight the effect on governance of the system being introduced in this project is a reflection of the goals of the proponents (and funders) of the system (the researchers, the International Development Research Centre and the government of Fez) rather than for example, the effect of the system on operational efficiency or effectiveness as might have been the case in other such implementations¹.

2 Governance

The concept of “Governance” has been found to have varying definitions. The World Bank Group’s [2] Poverty Reduction Strategy Paper (PRSP) defines Governance in terms of how “...power is exercised through a country’s economic, political, and social institutions”. In this sense, governance deals with the way political and socio-economic structures operate within a state.

The United Nations Development Program (UNDP) provides a more elaborated definition. Governance is

“the exercise of economic, political, and administrative authority to manage a country’s affairs at all levels. It comprises mechanisms, processes, and institutions through which citizens and groups articulate their interests, exercise their legal rights, meet their obligations, and mediate their differences” [2].

This definition indicates that governance has two major elements: a country and its citizens. Governance is how a political entity (i.e. state) organizes and administers its

¹ We are indebted to the anonymous reviewer for suggesting that this point be highlighted. It might be useful to note that it is in this area where the linkages can be most directly made between this project and the broad field of Community Informatics which is most directly concerned with the achievement of “social” and “community” goals through Information System design and implementation as well as or alongside more traditional Management Information System (MIS) goals of operational efficiency or effectiveness.

functions through exercising power via its various establishments. In addition, governance includes the sum of procedures, actions, and entities available to citizens (the people of a country) in order to enable them to conduct numerous operations, such as communicating their concerns, exercising their rights, undertaking their responsibilities, and arbitrating their disputes.

As a way of defining governance, the International Institute of Administrative Sciences (IIAS) differentiates between “government” and “governance.” Government is an institution that consists of a set of sub-institutions, namely “the constitution, legislature, executive and judiciary” [3]. In contrast, “governance” transcends government in that it “involves interaction between these formal institutions and those of civil society” [3]. In this sense, “governance” is the sum of relations taking place between the establishments of a government and those of civil society.

To further define “governance”, the IIAS identifies a number of components of which the concept is comprised; thus: “the degree of legitimacy, representativeness, popular accountability and efficiency with which public affairs are conducted” [4]. This suggests that there are four elements by means of which “governance” may be assessed: the presence of mechanisms that legitimize a system’s structures, processes, and actors; and the extent to which the system’s institutions and actors are representative, accountable to the general public, and effective.

Clearly, the definition of governance varies according to the institution. Each institution produces a definition of the term in question by highlighting one element or a set of elements that make up the nature of governance. Nevertheless, they have one element in common: governance is not synonymous with government. Rather, government is just one actor among several whose internal and external interactions with the other actors, such as civil society, the private sector and the general public, shape governance.

3 The E-Fez Project

The Fez E-Government (e-Fez) Project has as of March 2006, been in progress for one year with funding from the International Development Research Centre (IDRC) and conducted by a team of professionals at Al Akhawayn University in Ifrane Morocco. The project’s aim is to create a pilot E-Government system for the city of Fez that will facilitate citizens’ to receive governmental information and services easily, efficiently, quickly, and equally. As a background to this project, it was necessary to explain a variety of interlinked concepts in the area of governance, good governance, and E-Government. The intent of this was to identify ways to operationalize these broad notions in a manner capable of providing systematic and ongoing assessments of outcomes and ultimately impacts.

The project specifically is concerned with implementing an Information and Communication Technologies (ICT) system in support of Fez municipal service delivery in the ancient Moroccan capital of Fez and particularly services what is called the Bureau of “L’Etat Civil” (BEC). The BEC is the municipal government office which delivers many of the most widely requested services and document including Birth Certificates (BC). Although the BEC has daily and direct contact with local citizens it has in fact, remained to this time completely paper-based and

with uniformly manual process i.e. the systems are what we are terming “archaic”. These archaic systems present a variety of problems in delivering services to citizens.

It has been the intention of the e-Fez project to automate the BEC back office through digitization into a uniform and network-accessible database of all BEC Citizens’ Record Books². The project also developed and deployed an electronically enabled front office allowing for the automated processing of citizens’ birth certificates’ requests. This electronic “front office” provides three channels or modes for the submission and processing of a BC request: via a BEC employee desk, an interactive portal for the Fez community supported by an online service delivery platform, and via a related touch screen kiosk. The kiosk is available to the Fez local community at a pilot BEC site, free of charge, and usable by citizens, including those without literacy skills to facilitate automated submission and processing of BC requests.

Introducing ICT to a fully manual and paper-based service delivery is revolutionary in the Moroccan context. It has had the effect of automating the service delivery of Birth Certificates and has as a consequence generated numerous outcomes that have direct implications for the improvement of local governance. Aspects of the project are presented in other current or upcoming publications.

The Fez E-Government Project by automating the delivery of one of the most in demand citizen oriented services--birth certificates, will generate many outcomes that can be systematically evaluated in order to produce and disseminate knowledge on the deployment of such a project.

4 E-Fez Project Assessment

The need to assess and measure the project outcomes led us to the development and refinement of what we are terming an “Outcome Analysis” methodology. In this methodology we identified five major outcome categories that needed to be assessed: technology, organization, citizen, regulation, and good governance related outcomes. Concerning the project outcomes related to good governance, we proceeded in a systematic way. We began with the above identified UNDP definitions of the nine attributes of good governance. More specifically, we linked the e-Fez project to these attributes by developing working definitions adapted to the characteristics of the project.

Thus, in the context of e-Fez we developed the following project specific working definitions of the general UNDP attributes of good governance:

- **Transparency:** refers to bringing visibility to the service workflow for citizens by means of an automated service delivery
- **Effectiveness and efficiency:** refers to enabling optimal use of resources for citizens and tax payers in the service delivery

² The system being automated is one where all individual and family records are entered manually into large bound volumes (in triplicate) and kept in storage in the individual BEC for updating as required (through a change in an individual or family life-cycle event) or as might be required to authenticate the production of a necessary personal document (e.g. birth or marriage certificate).

- **Participation:** refers mainly to enabling the process of empowering citizens to legally control the service delivery to their advantage
- **Equity:** refers to citizens receiving the service on an equal basis
- **Rule of law:** refers to ensuring that the laws and regulations governing the service are applied in an impartial way
- **Accountability:** refers to creating standards against which the individuals providing the service and the service delivery can be held accountable
- **Responsiveness:** refers to serving all citizens in a consistent and predictable way
- **Consensus orientation:** is not applicable to this project which is concerned primarily with service delivery
- **Strategic vision:** is also not applicable to the project.

These working definitions identified for the project in turn enabled us to identify specific project outcomes and thus to track how the project is proceeding in relation to the broader definition and requirements of “good governance”—and in turn allow us to indicate the manner and degree to which the project was supporting the achievement of “good governance” within this particular service and institutional context.

5 Measuring Project Outcomes

From this basis and in close consultation with the stakeholders and the on-going development of the project we were able to identify the following project *outcomes* which in this instance can be directly linked to the previously identified *attributes of good governance*. The process here is one of reviewing the activities of the project, identifying outcomes and then determining through inspection which of these outcomes might be linked to which working definition. Alternatively this can be done by reviewing the working definition of each of the identified attributes of the overall definition to determine how the operations of, in, and through the project (the ICT installation) might have an indicative relationship to the specified attribute. Thus for example :

- **Transparency:** can be seen in a workflow that has become visible, transparent, and accessible for citizens via automated service delivery
- **Effectiveness and efficiency:** can be identified as when citizens and tax payers are enabled to have an optimal use of resources in the delivery and utilization of the public service
- **Participation/ empowerment:** occurs when citizens are empowered to legally control the service delivery to their advantage (thanks to the process of dis-intermediation that involves reducing/eliminating roles of middle people in service delivery)
- **Equity:** is realized when all citizens are served on an equal basis
- **Rule of law:** is achieved when laws and regulations are applied in an impartial way

- **Accountability:** arrives when standards are available against which individuals can be held accountable (thanks to the process of routinizing the service delivery; when the system is opaque, it is not possible to held individuals accountable)
- **Responsiveness:** is when all citizens are served in a consistent and predictable way.

The table below summarizes the project outcomes related to good governance:

Table 1. Summary of Project Outcomes Related to Good Governance:

UNDP attributes of Good Governance	UNDP Definition	Project working Definition	Project Outcomes
Transparency	“Transparency is built on the free flow of information. Processes, institutions and information are directly accessible to those concerned with them, and enough information is provided to understand and monitor them”	bringing visibility of workflow for citizens via automated service delivery	workflow become visible, transparent, and accessible for citizens via automated service delivery
Effectiveness and efficiency	“Processes and institutions produce results that meet needs while making the best use of resources”	enabling optimal use of resources for citizens and tax payers in service delivery	citizens and tax payers are enabled to have an optimal use of resources in receiving services
Participation	“All men and women should have a voice in decision-making, either directly or through legitimate intermediate institutions that represent their interests. Such broad participation is built on freedom of association and speech, as well as capacities to participate constructively”	process of empowering citizens to legally control the service delivery to their advantage	citizens become empowered to legally control the service delivery to their advantage (thanks to the process of dis-intermediation that involves reducing/eliminating roles of middle people in service delivery)
Equity	“All men and women have opportunities to improve or maintain their well-being”	serving citizens on an equal basis	all citizens are served on an equal basis
Accountability	“Decision- makers in government, the private sector and civil society organizations are accountable to the public, as well as to institutional stakeholders. This accountability differs depending on the organization and whether the decision is internal or external to an organization”	creating standards against which the individuals can be held accountable	standards are available against which the individuals can be held accountable (thanks to the process of routinizing the service delivery; when the system is opaque, it is not possible to held individuals accountable)
Responsiveness	“Institutions and processes try to serve all stakeholders”	serving all citizens in a consistent and predictable way	all citizens are served in a consistent and predictable way

Table 1. (continued)

Consensus orientation	“Good governance mediates differing interests to reach a broad consensus on what is in the best interest of the group and, where possible, on policies and procedures”	not applicable to the project dealing mainly with service delivery	not applicable to the project dealing mainly with service delivery
Strategic vision	“Leaders and the public have a broad and long-term perspective on good governance and human development, along with a sense of what is needed for such development. There is also an understanding of the historical, cultural and social complexities in which that perspective is grounded”	not applicable to the project dealing mainly with service delivery	not applicable to the project dealing mainly with service delivery

Having identified project outcomes, the team then reviewed the project activities to identify project outputs i.e. specific and measurable results from the project for which a quantitative or qualitative (indicative) measure could be assigned. The identified project outputs as linked to outcomes etc. can be captured from table 2 (colon 2).

6 The Outcome Analysis Methodology

The “Outcome Analysis” methodology provides a means by which those involved in the design, development and deployment of ICT systems (in this case e-governance systems) can assess the broader significance of those systems in relation to normative goals and using both qualitative and quantitative measurements. The logic of Outcome Analysis begins with formal (and generally accepted) definitions and their underlying characteristics (attributes)--including as in this case definitions of normative goals, moving through to project specific working definitions i.e. translating general definitions into the specific normative project goals (as identified by project stakeholders), further moving these to anticipated project outcomes and from there to specifiable (and measurable) project outputs.

Notably, this method can be applied to the range of normative goals, including goals in the variety of domains in which the implementation will have an effect i.e. technology, organization, service delivery, and so. It should be noted, that an “Outcome Analysis” approach is meant as a partial replacement for the more common concern in such projects with “Impact Assessment”³. The challenge with “Impact

³ As well to the more recent development of the Outcome Mapping Methodology cf. Sarah Earl, Fred Carden, and Terry Smutylo, “Outcome Mapping: Building Learning and Reflection into Development Programs”, IDRC 2001, http://www.idrc.ca/en/ev-9330-201-1-DO_TOPIC.html

Assessment” is that for most ICT implementations of the kind being discussed here the possible “impact” of the project will be very difficult to assess since “impacts” of ICT projects take a considerable period of time to emerge (impacts having to do with anticipated changes in the broader social and economic environment) and since the concern is with these broader changes, discerning these changes (and making any type of quantitative or even in most cases qualitative assessment) is extremely difficult.

Table 2. Attributes, Indicators and “Values”: Before and After Implementation

Governance Attributes	Measured Indicator	Value before automated system deployment	Value after automated system deployment
Transparency	Visibility of workflows for citizens via automated service delivery	<p>No</p> <p>Since the BEC back-office is completely manual, sub processes of making BC request, processing the request, and filling out the needed copies of BC are carried out in separated way (and sometimes with different employees). The citizen cannot monitor/ see the processing progress of his BC (e.g. the possibility of length/possible reasons for a delay in a processing are neither accessible nor visible)</p>	<p>Yes</p> <p>Since the BEC back-office is electronically enabled, sub processes of making BC request, processing the request, and printing the processed BC are merged in one process carried out on a real time basis. This secures the principle of: first-come-first-served</p>
Effectiveness and efficiency (as a citizen user)	Efficiency: optimal use of resources for citizens to request & obtain BC	<p>No</p> <p>requesting and obtaining BC is costly for citizens:</p> <ul style="list-style-type: none"> - extended waiting time - several trips to BEC - need to tip (or use social connections) 	<p>Yes</p> <p>Citizens making time/money/effort savings in requesting and obtaining BC:</p> <ul style="list-style-type: none"> - no waiting time - one trip to BEC - no trip
Effectiveness and efficiency (as tax payer)	Efficiency and effectiveness of using public scarce resources	<p>No</p> <p>To deliver BC, BEC needed 3 full time employees (when demand on BC is low and moderate). When demand on BC is high (during summer and early Fall period: from June to September): All BEC employees (10) stop processing their respective tasks in order to process BC requests Furthermore, they take BC requests home to be processed (which is illegal)</p>	<p>None: (casual calls on employee time with the elimination of full time 5 employees)</p> <p>No BC full time employee (any of the employee can instantly process BC requests while doing her other BEC related manual tasks)</p> <p>With the kiosk: no employee is needed to process the requests</p>

Table 2. (continued)

Equity	Citizens served with equity	<p>No</p> <p>Usually queuing/waiting creates motifs and conditions for bribery incidents. Citizens find themselves obliged to tip the employee in charge in order to be served, especially when they are in a hurry to meet tight deadlines of submitting paper work</p>	<p>Yes</p> <p>ICT eliminated the need for citizen to tip in order to be served</p> <p>all citizens are served on a timely and in a similarly professional manners (regardless of social class)</p>
Rule of law	Laws are applied impartially	<p>No</p> <p>Equity is violated; and violations are perceived as normal:</p> <p>Many violations of law as people paid for special privileges (queue jumping)</p>	<p>Yes</p> <p>Unnecessary need to tip reinforces the law of equity:</p> <p>Elimination of the need (opportunity) for violations of the law through tipping</p>
Participation/ empowerment	<p>- Citizens' active participation in BEC services</p> <p>- Dependency on bureaucracy: Dependence of citizens on the employees good will</p>	<p>No</p> <p>Citizens were not participating actively in the service delivery (with possible negative consequences on the service delivery arising from issues occurring in the workflow)</p> <p>Yes</p> <p>Citizens were at the mercy of employees to get served</p>	<p>Yes</p> <p>Citizens through the kiosk/online service delivery: Actively participate in the service delivery, which eliminates possibilities of negative consequences arising from difficulties in the workflow</p> <p>No</p> <p>Citizens through the kiosk/online service delivery: Are not at the mercy of employees</p>

In the “Outcome Analysis” approach which is being piloted in the eFcz project, rather than attempting to assess “impacts”, the concern is to assess shorter range and more specifiable “outcomes” (what the project is actually doing or producing) and for these, the development of indicators using the quite specific (and directly measurable) “outputs” of the project as in table 5 is quite feasible.

The above chart is given as an indication of the identified outputs used as indicators for the Project Outcomes and the method for assessing these and the identified value assigned to these in the context of the eFcz project.

7 Conclusion

The Outcome Analysis Project Assessment Method is an alternative method for assessing project results as a basis for linking these back into project goals and activities. The intention is to develop a method which includes both the integration of project stakeholder expectations and project goals into project assessment as well as to use such a process of project assessment as a means for planning, refining and adjusting a project's development as it is proceeding. This enables decision makers to have a clear understanding of the way the project and the system have contributed to overall development (in this case the contribution of the project to goals for enhanced "good governance" but these goals could cover the range of goals which are identified by the various project stakeholders in the range of Community Informatics implementations).

This method will be of particular interest to community based technology initiatives as a means for linking a collaboratively developed (community consensus) based project vision, with more specific technical project goals and then into assessable project outcomes. IT projects which have broader developmental or social goals (in addition to more traditional project goals of operational or administrative efficiency) would be a clear target for the application of this method.

References

1. Scholte, J. A.: The Globalization of World Politics. in Baylis, J. and Smith, S. (ed.): The Globalization of World Politics: An Introduction to International Relations, New York: Oxford University Press (2001)
2. The World Bank Group (WBG) 'What is Governance? Arriving at a common understanding of 'governance'', [Online], Available: <http://www1.worldbank.org/mena/governance/issues-keyQuestion.htm> [20 Aug 2005].
3. The Global Development Research Center (GDRC (b)) (2004) 'Understanding the Concept of Governance', [Online], Available: <http://www.gdrc.org/u-gov/governance-understand.html> [20 Oct 2005].
4. The Global Development Research Center (GDRC (a)) (2004) 'Governance: A Working Definition', [Online], Available: <http://www.gdrc.org/u-gov/work-def.html> [20 July 2005].

IS 2006 PC Co-chairs' Message

On behalf of the Program Committee of the 1st International Workshop on Information Security (IS 2006), it was our great pleasure to welcome the participants to IS 2006, held in conjunction with OnTheMove Federated Conferences (OTM 2006), from October 30 to November 1, 2006, in Montpellier, France. In recent years, significant advances in information security have been made throughout the world. The objective of the workshop was to promote information security-related research and development activities and to encourage communication between researchers and engineers throughout the world in this area.

In response to the call for papers, a total of 138 papers were submitted, from which 33 were carefully selected for presentation in 9 technical sessions. Each paper was peer reviewed by several members of the Program Committee or additional reviewers. The workshop program covered a variety of research topics, which are of current interest, such as multimedia security, network security, RFID security, cryptographic algorithms and protocols, biometrics for security, access control and smart card technology, risk analysis and business continuity, information security and service resilience. These technical presentations addressed the latest research results from international industry and academia and reported on findings on information security.

We thank all the authors who submitted valuable papers to the workshop. We are grateful to the members of the Program Committee and to the additional reviewers. Without their support, the organization of such a high-quality workshop program would not have been possible. We are also indebted to many individuals and organizations that made this event happen, namely Springer. Last but not least, we are grateful to the General Co-chairs, Workshops General Chair, and the Web Site Manager for their help in all aspects of the organization of this workshop.

We hope that you enjoyed this 1st International Workshop on Information Security at Montpellier, France, if you attended, and that you found it a useful forum for the exchange of ideas, results and recent findings.

August 2006

Mário Freire, University of Beira Interior, Portugal
Simão Melo de Sousa, University of Beira Interior, Portugal
Vitor Santos, Microsoft Lisbon, Portugal

An Implementation of a Trusted and Secure DRM Architecture

Víctor Torres, Jaime Delgado, and Silvia Llorente

Universitat Pompeu Fabra, Passeig de Circumval·lació, 8,
08003 Barcelona, Spain
{victor.torres, jaime.delgado, silvia.llorente}@upf.edu
<http://dmag.upf.edu>

Abstract. Content providers and distributors need to have secured and trusted systems for the distribution of multimedia content with Digital Rights Management (DRM) to ensure the revenues derived from their works. This paper discusses the security mechanisms applied to the implementation of a DRM architecture, regarding the certification and verification of user tools integrity during their whole life cycle, the mechanisms for providing a secure and trusted communication between client tools and the server framework for authorisation, certification or verification purposes, and the mechanisms for the secure storage and resynchronisation of the reports that describe the actions performed by users during the tool offline operation. The presented architecture is being implemented in the AXMEDIS project, which aims to create an innovative technology framework for the automatic production, protection and distribution of digital cross media contents over a range of different media channels, including PC (on the Internet), PDA, kiosks, mobile phones and i-TV.

Keywords: Secure content management, multimedia content protection, digital rights management systems.

1 Introduction

In [1] [2] [3] we presented in a general way an architecture to manage multimedia information taking into account digital rights management (DRM) and protection. The architecture, called DMAG Multimedia Information Protection and Management System (DMAG-MIPAMS), whose name is after our group acronym DMAG [4], consists of several modules or services, where each of them provides a subset of the whole system functionality needed for managing and protecting multimedia content. The architecture is depicted in Figure 1.

In this paper we are going to give more details about a real implementation of that architecture which is being developed in the context of the AXMEDIS European Project [5]. In particular, we will concentrate on how communications and services in the architecture can be secured and trusted, and which mechanisms have been introduced to ensure that client tools act as expected and are not modified by malicious users.

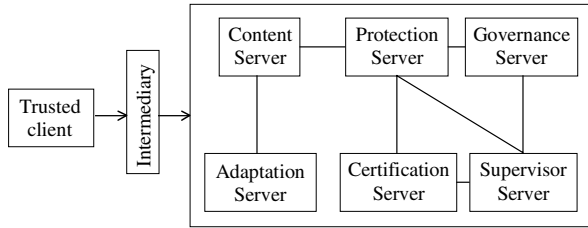


Fig. 1. DMAG MIPAMS architecture

In next sections we provide an overview of the AXMEDIS architecture. Then, we analyse the trust and security aspects and finally we provide a use case to understand how it works in a real scenario.

2 An Implementation of the Architecture

The architecture implemented in the AXMEDIS project consists on several independent modules that interact as web services when they are located in different machines or directly in other situations.

The general description of the AXMEDIS architecture main modules, depicted in Figure 2, is as follows:

- *Protection Processor*. This client tool module is responsible for estimating the client tool fingerprint, enabling or disabling the tool, verifying the tool integrity and unprotecting protected multimedia objects.
- *Protection Manager Support Client (PMS Client)*. This client tool module manages and stores protection information, licenses, reports regarding the offline performed actions and other secured information in a local secure storage system called secure cache. It is responsible for authorising users to perform actions over objects with respect to digital licenses during offline operation. It also delivers protection information to the protection processor, if present in the secure cache, or requests it to the AXCS after a positive authorisation. It acts also as the intermediary module used by Protection Processor to contact AXCS to certify and verify tools.
- *Protection Manager Support Server (PMS Server)*. This server side module is responsible for authorising users to perform actions over objects in an online environment and requesting protection information to the AXCS if needed. It acts also as an intermediary module to contact AXCS from PMS Client.
- *AXMEDIS Certifier and Supervisor (AXCS)*. AXCS is the authority in charge of user and tool registration (Registration Web service), user and tool certification (AXMEDIS Certification and Verification, AXCV), user and tool management (e.g. status supervision, automatic blocking, deadline supervision, etc.), user and tool unique identifier generation and object metadata collection. AXCS is also responsible for saving the Protection Information related to protected multimedia objects as well as the actions performed on them (AXMEDIS Supervisor, AXS), the so-called Action Logs. Action Logs are the particular implementation of MPEG-21 [6] Event Reports [7] in the AXMEDIS context. AXCS also includes a

user Registration service, useful for registering new users in the system from distribution servers. All these data are stored in the AXCS database, which is accessed through the AXCS database interface module in order to keep the access independent from its implementation. Other functionalities provided by AXCS are those related to reporting and statistical analysis, which are performed by the Core Accounting Manager and Reporting Tool (CAMART module) by analysing the information stored in the AXCS database and collected in Action Logs. The integral modules of AXCS (see Figure 2) have been developed as web services or libraries.

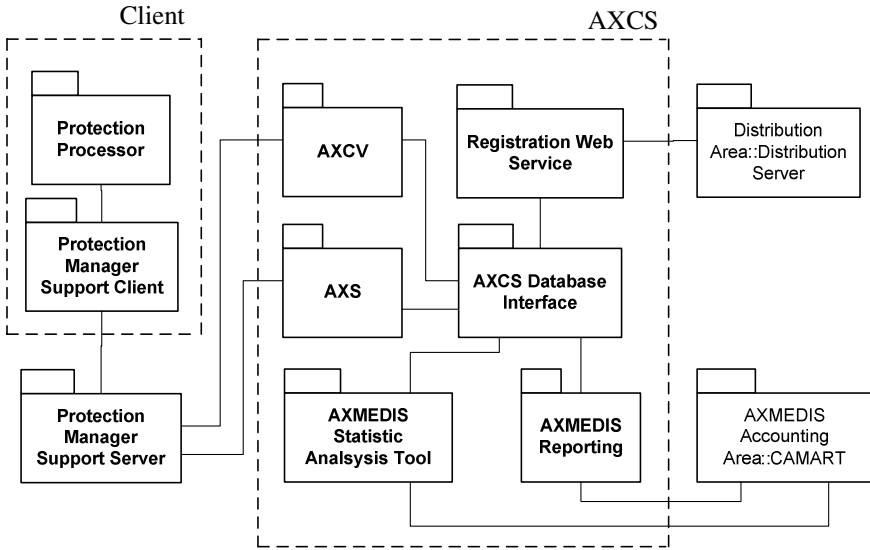


Fig. 2. AXMEDIS architecture regarding protection, rights management and accounting functionalities

2.1 Security and Trust on User Tools and Communication to Server

As we have mentioned in previous sections, in the client side we have different modules as Protection Processor and PMS Client which are devised to communicate with the server part by providing not only security to the transactions but also trust from the server side perspective. In the following sections we are going to describe the different mechanisms that the system includes to achieve the security and trust goals.

2.1.1 Registration of Users

All the users in the system must register, which enables their interaction with the system and system tools. User information is stored in the server side (AXCS) and is used for further verification purposes. After the user registration, the corresponding AXCS issues a user certificate that will be used to authenticate the user when performing some specific operations as the certification of tools (see section 2.1.2).

Every AXMEDIS user has associated a status that is used to determine whether the user is blocked or not in the system when interacting with the server part. The user status can be modified by the AXCS if some critical operation attempt is detected.

2.1.2 Registration of Tools

Tools using AXMEDIS framework must be verified to accomplish a series of guidelines, which are checked before registration is done. Once verified, each tool is registered for being used by AXMEDIS users. During registration phase, a fingerprint of the software tool is estimated so that its integrity can be checked later when the tool is installed and certified or verified on a specific device, as we will see in next sections.

2.1.3 Certification of Tools

The certification of a tool that uses AXMEDIS framework is a necessary step for that tool to work. Before a user is able to run and use a tool, the tool must connect to the AXCS to be certified as an “installed tool”. Before installation, AXCS verifies the tool integrity by comparing its fingerprint to the one stored during the tool registration process and, once installed, extracts some information (tool fingerprint) concerning the installation of the tool and the device where it is installed.

A malicious user who tries to certify a tool whose fingerprint does not match the original registered tool fingerprint would be automatically blocked in the system so that he cannot continue performing other operations within the system. Moreover the tool would not be certified and thus it would not be operative.

Once a user successfully certifies a tool, any user of the system who owns a valid AXMEDIS user certificate can use it. Blocked users cannot use tools in the system.

To perform the certification of a tool, the tool connects to the AXCS via PMS Client and PMS Server web service. In order to have a secure communication, the client certificate is used to authenticate the user against the PMS Server.

The certification process involves different operations in the AXCS:

- *Generation of tool certificate and private key.* AXCS Certification Authority generates a tool certificate. It is used to establish secure communications, via SSL providing secure web services, to the PMS Server by any user that manages the certified tool. In this way we ensure that only certified tools can interact with the server part in an authenticated manner.
- *Generation of tool unique identifier.* A tool unique identifier is assigned to that specific installation of the tool and is used to identify it when interacting with the server side. The identifier is generated following the UUID format [8] and inserted in the tool certificate.
- *Generation of tool activation code.* A tool activation code is used to enable the tool operation. Some cryptographic algorithms that depend on the specific installation are used to generate it and they are inserted in the tool certificate as a certificate extension.
- *Generation of tool fingerprint.* The tool fingerprint, as we have already said, concerns the installation and the device where the tool is installed. This fingerprint is used in further verification process to determine if the tool has been manipulated

or if the device has changed or, in other words, to ensure the tool is still trusted in further executions.

- *Storage of identifier, activation code, tool fingerprint and certificate.* All the previous information is stored in the AXCS database and will be used to authenticate the tools that connect to the server part and to verify their integrity, as we will explain in next sections.

On the other hand, the certification process supposes also different operations in the client side (PMS Client and Protection Processor):

- *Reception of tool certificate, private key, tool identifier and activation code.* Regarding the tool certificate, private key, tool identifier and tool activation code, tool identifier and tool activation code are included in the tool certificate in the following manner (see Figure 3): 1) The tool unique identifier is used as the certificate common name (CN) in the subject distinguished name (DN) field; 2) The tool activation code is inserted as a certificate extension.

```

Data:
  Version: 3 (0x2)
  Serial Number: 1000000493 (0x3b9aced)
  Signature Algorithm: sha1WithRSAEncryption
  Issuer: O=AXMEDIS, OU=AXMEDIS AXCS CA, C=ES, CN=AXMEDIS
  AXCS CA/emailAddress=axmedis@axmedis.org
  Validity
    Not Before: ...
    Not After: ...
  Subject: O=AXMEDIS, CN=ITO_cdec4a1-dbc-362c-a30d-bb936342996c
  Subject Public Key Info:
    Public Key Algorithm: rsaEncryption
    RSA Public Key: (1024 bit)
      Modulus (1024 bit): ...
      Exponent: 65537 (0x10001)
  X509v3 extensions:
    X509v3 Subject Key Identifier: ...
    X509v3 Authority Key Identifier: ...
    1.3.6.1.4.1.25576.1.1.1: ...
  Signature Algorithm: sha1WithRSAEncryption
  ...

```

Fig. 3. AXMEDIS tool certificate fields

The tool activation code extension is identified with the Object Identifier 1.3.6.1.4.1.25576.1.1, where 1.3.6.1.4.1.25576 is the Private Enterprise Number assigned by IANA to AXMEDIS Organisation and 1.3.6.1.4.1 corresponds to IANA-registered Private Enterprises [9] (see Figure 4).

Current assignation of the AXMEDIS tree corresponding to the 1.3.6.1.4.1.25576 branch is the following:

1.3.6.1.4.1. 25576.0: reserved
1.3.6.1.4.1. 25576.1: AXMEDIS PKI-X.509 related objects
1.3.6.1.4.1. 25576.1.1: AXMEDIS Tool certificate extensions
1.3.6.1.4.1. 25576.1.1.1: AXMEDIS Tool activation code (or enabling code)

Fig. 4. Assignment tree corresponding to the AXMEDIS IANA Enterprise number

The tool certificate and private key are finally packaged by AXCS in a PKCS12 [10] structure protected with a password linked to the user that performed the certification and delivered over the secure channel established using the user and server certificates.

- *Storage of certificate and private key and tool activation.* The PKCS12 structure is accessed by Protection Processor in order to extract the tool certificate and private key, which are finally stored in a local keystore, and also to get the activation code used to enable the tool.

2.1.4 Secure Communication

As we have already mentioned, all communications between client tools and the server part are performed over a secure channel, which is established by means of client and server certificates, thus having authentication of both parties. Whereas before client tool certification client tools need to use user certificates, after certification they use tool certificates to create the secure channel with PMS Server. PMS Server also establishes a secure communication with AXCS by means of its own server certificate issued by the AXCS CA. It is worth noting that the certificates issued to users, tools and servers have different certificate purposes.

2.1.5 Verification of Tools

Verification of tools is devised to cover two functionalities. First, it provides a means to ensure that client tools have neither been manipulated nor corrupted. Moreover, verification is used to resynchronise all the actions performed by users during offline operation, that were stored in the local secure cache.

Verification of tools is performed periodically by the Protection Processor and every time the user tool resynchronises the offline performed actions with the server part. It consists on the verification of the estimated tool fingerprint in the moment of the verification against the tool fingerprint stored in AXCS database during the certification of the installed tool.

Regarding the tool integrity verification, if AXCS detects that critical parts of the tool or the device have been manipulated, it can adopt the pertinent measures as, for example, blocking the specific installed tool for which the verification failed.

Regarding the resynchronisation of offline performed operations, AXCS executes an algorithm to determine whether the received list of operations, which are called Action Logs in the AXMEDIS context, is complete with respect to the previous received operations. This integrity check is feasible thanks to the calculation of a fingerprint on the performed Action Logs, which is computed by PMS Client during the tool operation. This fingerprint is sent to AXCV when resynchronising the offline Action Logs and is verified by AXCV using the algorithm depicted in Figure 5.

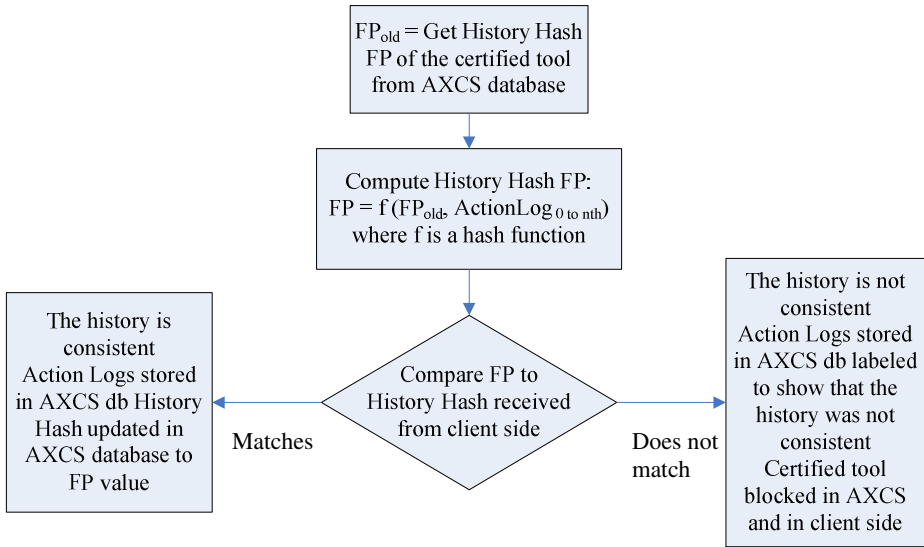


Fig. 5. Algorithm to determine the integrity of the received Action Log list in AXCV

The history fingerprint (FP) computation is also performed in the client side for each action performed in the online or offline operation, so that, once synchronised it must hold the same value in both PMS Client and AXCS. When an online operation is performed, this value is immediately synchronised in the server side. When any offline actions are performed, an Action Log associated to each of them is stored in the local secure cache, where PMS Client computes and separately stores FP value after the operation.

3 Use Case

In this section we present a scenario to illustrate how the proposed architecture and the processing of protected and governed multimedia content are related. It describes content consumption.

Imagine that a user has purchased online a license that grants him the right to play a movie during a certain period of time. The acquisition of the license could be performed in various ways. On one hand, the user could have obtained the license in the same place where he purchased the content. In this case, if the license needs to be customised for a particular user, the content distributor must request the license to the corresponding protection and governance servers. On the other hand, the user could have obtained the content through a P2P network, or other online or even offline distribution channels. In this latter case, the content must have some metadata that identifies the content server with which the user must interact to purchase the appropriate license.

The aforementioned user has, installed in his device, a specific tool or plug-in that manages the protected and governed objects of the proposed system and that is able to display them in the appropriate way.

The use case begins when the user downloads a protected and governed movie, opens it with his favourite player, which includes the appropriate plug-in and tries to watch it (Play movie). Although the plug-in has not been manipulated, the system needs to verify its integrity and certify it before allowing its operation.

Figure 6 shows the steps involved in the content consumption use case, which are the following:

1. The viewer requires unprotecting the movie to an internal trusted module, the Protection Processor.
- 2-3. Protection Processor estimates the installed tool fingerprint and connects to AXCS through PMS Client and Server in order to certify the tool.
- 4-5. AXCS successfully verifies user data and status and tool integrity with respect to registered tool.
6. AXCS sends Protection Processor a PKCS12 structure that contains tool private key and tool certificate with the tool identifier and activation code.
- 7-8. Protection processor stores tool certificate and private key in a local repository, extracts activation code and enables tool operation.
- 9-11. Before the authorisation, Protection Processor always calls verify method to check tool integrity and resynchronise offline Action Logs. In order to call it, it must reestimate the tool fingerprint and extract user and tool information from pertinent certificates.
12. PMS Client gets action logs from secure cache and contacts AXCS through PMS Server. (Note that in this case, as it is the first usage, there will not be any action logs)
- 13-17. AXCS verifies user and tool data with respect to certified tool Fingerprint, computes and verifies the operation History Fingerprint and stores received action logs in the AXCS database.
18. The result of the verification is sent to Protection Processor.
19. Protection Processor asks for authorisation and for protection information to PMS Client. As the user is working online, PMS Client contacts PMS Server.
- 20-21. PMS Server contacts AXCS to retrieve the object protection information.
22. PMS Server performs the license-based authorisation using its license repository.
- 23-24. As the authorisation is positive, PMS Server sends the pertinent Action Log to AXCS, which stores it in its database.
25. PMS Server notifies PMS Client the successful authorisation
- 26-27. PMS Client updates and stores the operation history hash fingerprint and the object protection information in the local secure cache.
28. PMS Client notifies Protection Processor the successful authorisation
- 29-31. Protection Processor requests the Protection Information to PMS Client, which retrieves it from the local secure cache.
- 32-33. Protection processor is capable of unprotecting the protected object so that the player can finally display the film to the user.

It is worth noting that, once the tool is certified, only verification process is done when the user wants to consume multimedia content. Steps 2 to 10 are no more executed after tool certification.

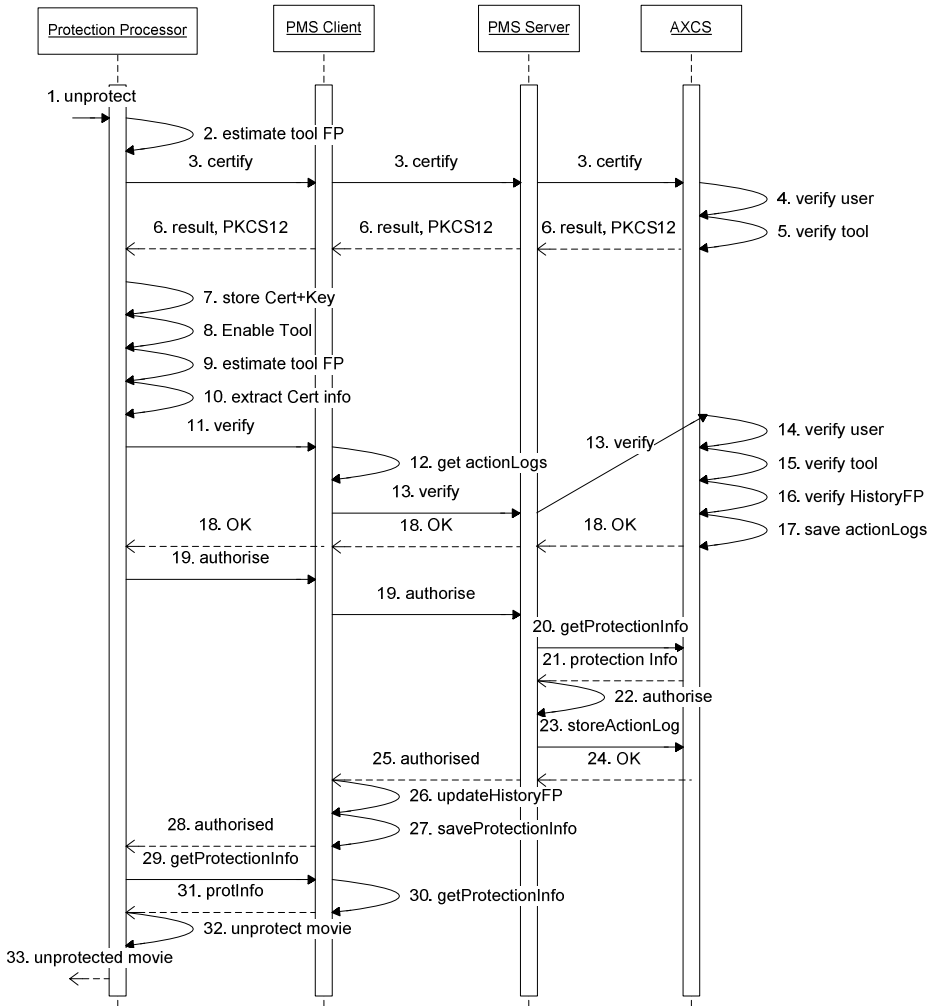


Fig. 6. Use Case

4 Conclusions

In this paper we have presented a possible implementation of the DRM architecture presented in [1] [2] [3], which is being developed in the context of the AXMEDIS European project [5]. In particular, we have concentrated in the aspects that provide security and trust to the interaction between user tools and the server part of the system, such as the registration of users and tools, the certification of tools, the establishment of secure channel communications using both client and server authentication and the verification of tools during their whole life operation. We have also provided a use case to illustrate the whole content consumption process.

Several demonstrators over different distribution channels (Satellite, PC, Kiosk, etc.) have been produced within the AXMEDIS project in order to validate the proposed solution and show its potential usage. Moreover, a public framework will be provided for the adoption of the AXMEDIS solution. Demonstrations of single tools and also of the framework are provided at AXMEDIS conferences and sometimes on the AXMEDIS portal [5]. The framework can be accessed by all affiliated partners.

The next steps to be tackled involve the integration with other existing content production and distribution tools in order to facilitate interoperability of both content management systems and multimedia and cross media protected objects.

Acknowledgements. This work has been partly supported by the Spanish administration (DRM-MM project, TSI 2005-05277) and is being developed within AXMEDIS [5], a European Integrated Project funded under the European Commission IST FP6 program. A special mention should be done to DSI-DISIT [11] for their collaboration in the work presented in this paper.

References

1. Torres, V., Rodríguez, E., Llorente, S., Delgado, J.: Trust and Rights in Multimedia Content Management Systems. Proceedings of the IASTED International Conference on Web Technologies, Applications, and Services (WTAS 2005). ACTA Press, Anaheim Calgary Zurich (2005) 89-94
2. Torres, V., Rodríguez, E., Llorente, S., Delgado, J.: Use of standards for implementing a Multimedia Information Protection and Management System. Automated Production of Cross Media Content for Multi-Channel Distribution (AXMEDIS 2005). First International Conference on. IEEE Computer Society, Los Alamitos Washington Brussels Tokyo (2005) 197-204
3. Delgado, J., Torres, V., Llorente, S., Rodríguez, E.: Rights and Trust in Multimedia Information Management. 9th IFIP TC-6 TC-11 Conference on Communications and Multimedia Security (CMS 2005). Lecture Notes in Computer Science, Vol. 3677. Springer-Verlag, Berlin Heidelberg New York (2005) 55-64
4. Distributed Multimedia Applications Group (DMAG), <http://dmag.upf.edu>
5. Automatic Production of Cross Media Content for Multi channel Distribution (AXMEDIS), IST 2004 511299, <http://www.axmedis.org/>
6. MPEG 21, <http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm>
7. ISO/IEC, ISO/IEC FDIS 21000-15 – Event Reporting
8. A Universally Unique Identifier (UUID) URN Namespace, <http://tools.ietf.org/html/4122>
9. Internet Assigned Numbers Entity (IANA) Private Enterprise Number, <http://www.iana.org/cgi-bin/enterprise.pl>
10. PKCS #12 v1.0: Personal Information Exchange Syntax Standard. RSA Laboratories, June 24, 1999, <http://www.rsasecurity.com/>
11. Distributed Systems and Internet Technology Lab - Department of Systems and Informatics DSI-DISIT, University of Florence, <http://www.disit.dsi.unifi.it/>

Using Image Steganography for Decryptor Distribution

T. Morkel¹, J.H.P. Eloff², and M.S. Olivier³

Information and Computer Security Architecture (ICSA) Research Group
Department of Computer Science, University of Pretoria, 0002, Pretoria, South Africa
{tmorkel, eloff, molivier}@cs.up.ac.za

Abstract. When communicating secret information there is more than one route to follow to ensure the confidentiality of the message being transmitted. Encryption might be an obvious choice; however there are limitations and disadvantages to using encryption. An alternative approach is steganography, which is a technology for hiding information in other information. Combining the two disciplines may provide better security but more overhead, since the receiver must now have knowledge not only of how the information was encrypted, but also of how it was hidden, in order to retrieve the message. This paper proposes a system where image steganography is combined with encryption by hiding not only a message inside an image, but also the means to extract and decrypt the message. An executable program is hidden inside the image that functions as a decryptor, enabling the receiver to be oblivious to the encryption algorithm used.

1 Introduction

Encryption enables private communication by scrambling a message in such a way that it can only be recovered by using an appropriate decryption program in combination with an appropriate key. Encryption, however, suffers from a number of drawbacks – notably the fact that the mere presence of an encrypted message might be cause for suspicion.

Another drawback of encryption is the limitations that have been enforced by certain governments [1]. The use of encryption – and even the possession of an encryption algorithm – is illegal for ordinary citizens in some countries. This often implies that a traveler has to delete any encryption software when entering a country and is only allowed to acquire and install it again after leaving that country. Additional issues of encryption often imply that the receiver needs a number of decryptors and may have to occasionally get rid of them and reinstall them. People who wish to communicate in secret must thus find alternative ways of doing so.

Steganography, a technology used for hiding information in other information [2], is one such way. While steganography and encryption have their separate drawbacks, combining them result in a system that builds on the benefits of both. By first encrypting information and then embedding it inside an image, steganography adds another layer of security to encryption. An eavesdropper will first need to identify the embedded information, then extract the information and then finally decrypt it to use the secret.

Unfortunately, there is a drawback to this combination, namely the amount of overhead. With single encryption, as with single steganography, the receiver only has to have knowledge of the encryption, or steganographic, algorithm used to obtain the message. However when combining encryption and steganography, the receiver needs to not only have knowledge of how to decrypt the information, but also of how to extract it. This brings us to a problem similar to the cryptographic software distribution problem, where the software needed to decrypt the message has to be communicated to the receiver along with the encrypted message, making it harder to ensure the confidentiality of both.

This paper presents a solution to this problem by not only embedding the encrypted message in the image, but to embed the software to decrypt the message along with it, using steganography to distribute the decryptor on demand.

The remainder of the paper is structured as follows: Section 2 provides the reader with a brief overview of image steganography since it is a lesser known technology than encryption. Section 3 looks at the proposed design of the system. In Section 4 the advantages and the potential weaknesses of the proposed system are discussed and in Section 5 a conclusion is reached.

2 Overview of Steganography

Although many different cover mediums can be used for embedding information, images are the most popular mainly because of the redundancy created in the way that digital images are stored. In an environment where the Internet is used, images are also a common multimedia format, making it an ideal carrier for information, while reducing suspicion.

Image steganography techniques can be divided into two groups: those in the Image domain and those in the Transform domain [3]. Image – also known as spatial – domain techniques embed information in the intensity of the pixels directly and encompass bit-wise methods that apply bit insertion and noise manipulation.

For the Transform – also known as frequency – domain, images are first transformed, then the message is embedded in the image and they involve the manipulation of algorithms and image transforms [4]. These methods hide information in more significant areas of the cover image, making it more robust [5].

The simplest and most common technique for embedding information in images is called the Least Significant Bit (LSB) technique [6]. The least significant bit (the 8th bit) of some or all of the bytes inside an image is replaced with a bit from the secret message. When using a 24-bit image, a bit of each of the red, green and blue colour components can be used, since they are each represented as a byte. In other words, one can store 3 bits in each pixel. For example, 3 pixels from a 24-bit image can be as follows:

```
(00101101  00011100  11011100)
(10100110  11000100  00001100)
(11010010  10101101  01100011)
```

When the number 200, which binary representation is 11001000, is embedded into the least significant bits of this part of the image, the resulting pixels are as follows:

```

(00101101  00011101      11011100)
(10100110  11000101      00001100)
(11010010  10101100      01100011)
    
```

Although the number was embedded into the first 8 bytes of the grid, only the 3 underlined bits needed to be changed according to the embedded message. On average, only half of the bits in an image will need to be modified to hide a secret message using the maximum cover size [7]. Since there are 256 possible intensities of each primary colour, changing the LSB of a pixel results in small changes in the intensity of the colours. These changes cannot be perceived by the human eye - thus the message is successfully hidden.

3 System Design

The basic idea behind the proposed approach is to use steganography as a means of communicating secret, encrypted information along with the decryptor program.

3.1 System Specification

The system is divided into two phases: the embedding phase and the extracting phase.

Embedding Phase. The embedding phase is responsible for encrypting the secret message and embedding it into the image. Although any steganographic algorithm can be used, for the purposes of this research LSB will be used as an example together with the bitmap (.BMP) image file format. Knowledge of the encryption algorithm used is not imperative at this stage of the discussion.

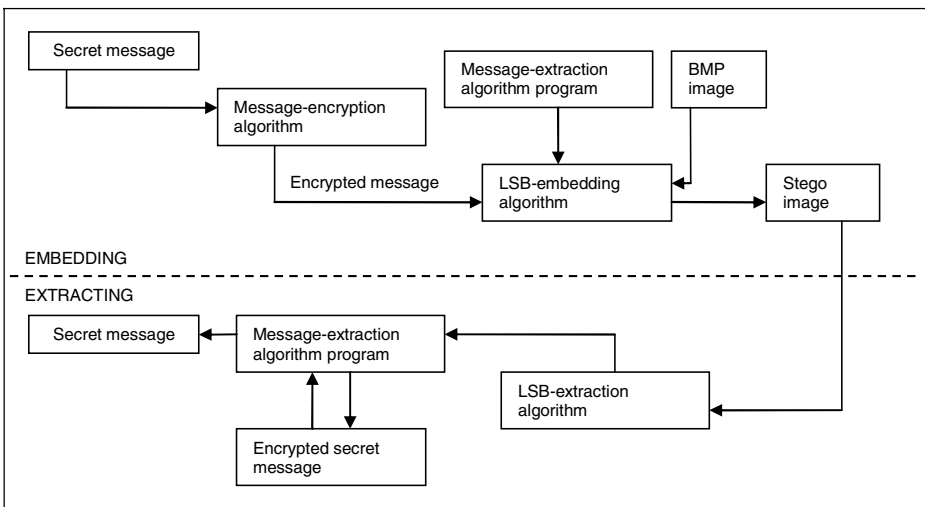


Fig. 1. System Design

The system consists of four algorithms. These algorithms are shown in Fig. 1. The first algorithm, the message-encryption algorithm, is simply used for encrypting the message and depends entirely on the encryption algorithm used.

The message-extraction algorithm is the algorithm used to extract and decrypt the message at the receiver's end. This algorithm is not explicitly used in the embedding phase, but has to be embedded into the image along with the message. This algorithm can either be in source code or converted to an executable (.EXE) program. There are advantages and disadvantages to both approaches that will be discussed later on in the paper.

The third algorithm, the LSB-embedding algorithm is used to hide the encrypted message along with the message-extraction algorithm. An inverted version of the LSB-embedding algorithm, the LSB-extraction algorithm, has to be communicated to the receiver through secure channels.

A specific format is specified for embedding information in the LSB-embedding algorithm.

```
<filename.extension>$<4 bytes program size embedded in 32 bytes>message-extraction algorithm program
```

- The first 52 bytes of a BMP image consists of header data and cannot be modified
- The receiver will need to execute this program and since he will have no knowledge of the type of file beforehand, his ability to do so will directly depend on the inclusion of the filename and format. For this reason, after the first 52 bytes of the image, the filename of the embedded program, with special regard to the file extension, is hidden. Using LSB the embedding of the filename will start in the 53rd byte and continue in the following sequential bytes.
- The filename is followed by a dollar sign to indicate the end of the filename
- Since the receiver is unaware of the type of information to expect, he will also have no knowledge of the amount of information to extract. For this reason 4 bytes are used for storing the size of the message-extraction algorithm. If an attempt is made to extract more information than is actually embedded, the embedded program will not be able to execute accurately.
- Finally the message-extraction algorithm is embedded

LSB-embedding-algorithm. Let I be the carrier image with I' the image converted into

binary. Each pixel in I is denoted as I'_i with i the pixel number. Each pixel consists of three colour components denoted as $I'_{i,RED}$, $I'_{i,GREEN}$ and $I'_{i,BLUE}$.

Let S be the secret message, converted into binary, S' , and encrypted using the message-encryption algorithm resulting in $E(S')$.

Let P be the message-extraction algorithm program, converted into binary, P' with each bit denoted as P'_x where x is the bit number.

Let N be the filename of the message-extraction algorithm program, converted into binary N' with each bit denoted as N'_y , where y is the bit number

Calculate the size of the program P' in bits, denoted as F and converted into binary, F'


```

Set the value of  $i$  to 53 and For each bit in  $N'$ 
Replace the LSB of the next pixel's  $I'_i$  three bytes as follows
  Replace the LSB of  $I'_{i,RED}$  with a bit from  $N'$ 
  Replace the LSB of  $I'_{i,GREEN}$  with a following bit from  $N'$ 
  Replace the LSB of  $I'_{i,BLUE}$  with a following bit from  $N'$ 
  Increment  $i$ 
Convert the $ sign into binary,  $D'$ 
For the next 8 bytes of  $I'$ 
  Replace the LSB of the next byte in  $I'$  with a bit from  $D'$ 
For the next 32 bytes of  $I'$ 
  Replace the LSB of the next byte in  $I'$  with a bit from  $F'$ 
While not the end-of-file of  $P'$ 
  Replace the LSB of the next byte in  $I'$  with a bit from  $P'$ 
While not the end-of-file of  $E(S')$ 
.
.
.
          (depends on the manner in which  $E(S')$  is hidden)
.
.

```

After the message-extraction algorithm is embedded, the message can be embedded in a number of different ways. The least secure way would be to continue from the last byte of the message-extraction algorithm, since it would make the message easier to locate in the event of discovery of the decryptor program. If someone were to uncover the program, it would not necessary mean that they would suspect that there is more information embedded in the image. It would thus be wiser to use a different method to embed the secret message. One can either start embedding from the end of the image, or use selective LSB and only use a predetermined sequence of bytes. There are more possibilities that can be explored.

Extracting Phase. At the receiver's end, the extracting phase is where the program is extracted and executed to extract and decrypt the message. Two of the four algorithms are used in the extracting phase.

Using the LSB-extraction algorithm obtained from the sender, the message-extraction algorithm is retrieved from the image and stored in the appropriate file. Depending on whether the program is in source code or an executable program, the program can either be compiled and executed or simply executed. The program will receive the communicated image as input, locate and extract the message bits, and decrypt it.

LSB-extraction algorithm. Using the same definitions as the LSB-embedding algorithm, the following:

```

Set the value of  $i$  to 53
While  $N_y$  is not the $ character
  Read in the LSBs of 8 bytes of  $I'$  at a time
  Convert the bits into ASCII and store in  $N_y$ 
For the next 32 bytes

```

```

    Read in the LSBs of a byte of  $I'$ , store it in  $F'$  and convert it into an integer
    number  $F$ 
while  $F \geq 0$  do
    Read in the LSBs of 8 bytes of  $I'$  at a time
    Convert the bits into ASCII and store in  $P_x$ 
    Save  $P$  in a file called  $N$ 

```

3.2 System Considerations

The efficiency and functionality of the system can be measured with regards to invisibility and payload capacity. Invisibility being the first and foremost requirement since the strength of a steganographic system lies in its ability to be unnoticed by the human eye [8]. The invisibility of the system will depend entirely on the steganographic algorithm used, since some algorithms are more successful at unperceivable data hiding than others.

Payload capacity is the amount of data that can be stored in an image [8]. This is especially important in deciding which encryption algorithm to implement, since a very complex algorithm would result in a large program file that would not be able to fit into an image of reasonable size.

3.3 Prototype construction

Several prototypes were developed to implement the proposed system. Usually simple encryption algorithms were used, since the prototypes were developed to test the feasibility of implementing the proposed system and not the strength of the encryption. A

comparison of two example prototype implementation is given in Table 1.

Table 1. Comparison of prototype implementations

	Embedded program	Embedded program function	Message size	Encryption	Embedded program size	Minimum image size	Payload capacity
Project A	Java class	Used for message encryption and decryption	150 bytes	Built-in DES function	4.6 KB	118 KB (200 x 200 pixels)	4%
Project B	Java class	Used for message extraction and decryption	150 bytes	Permutation and XOR	2.8 KB	30 KB (100 x 100 pixels)	9%

3.4 Source Code or Executable Program?

Whether the embedded code has to be source code or an executable program, will depend on a number of factors. The advantages and disadvantages for use in the proposed system will need to be investigated.

When using source code, the receiver will be able to examine the embedded program. This could be useful when the sender wishes to communicate not only the message but also a specific encryption algorithm that he might want to use for future communications. This would mean that the sender need not send the decryptor program to a specific party in every following communication.

Another advantage for the receiver being able to examine the source code before executing it, is that there is always the possibility that the program in the image originates from a malevolent sender that might embed a malicious program in the image. When the program is an executable the receiver has no option but to blindly execute the program, having no idea what the program will do to his computer. The solution to this is a trust issue and will amount to the receiver trusting that the sender has not included any code that might damage his computer system. Alternatively a similar approach to a Java sandbox can be used to ensure that executable code does not gain access to the receiver's computer resources [9].

Depending on the programming language and compiler used, it is a very complicated task to decompile an executable program, in other words to retrieve the original source code from an executable program [10]. This can be made even more difficult when using code obfuscation, which is a technique for deliberately making source code difficult to read [11]. Should the nature of the sender/receiver relationship call for the confidentiality of the encryption algorithm itself, an executable program would be more suitable. All the receiver can do is to execute the program and receive the decrypted message, without being able to gain knowledge of how the message was encrypted or decrypted.

The risk of discovery also plays a role in deciding whether to embed source code or an executable program in the image. Due to its nature, executable code gives the impression of being more like random data than source code, making it more difficult to notice should someone be looking for hidden information. Source code, being a close resemblance to natural language, is more prone to statistical attacks.

Finally an advantage of using an executable program over using source code is concerned with platform independence. An executable program will be able to execute on any platform, while some platforms might not be able to compile and execute certain source code. Along the same lines, communicating source code to a receiver is based on the assumption that the receiver actually possesses the correct compiler software to compile and execute the source code. This might not always be the case.

4 Advantages and Potential Weaknesses of the Proposed System

The concept of combining encryption with steganography in such a way to hide not only an encrypted message in an image, but also the decryptor program, holds many advantages over other forms of secret information communication. Unfortunately there are also potential weaknesses to the system.

4.1 Advantages of the Proposed System

The main advantage that the proposed system offers is by combining encryption and steganography you also combine their individual benefits. Cryptography mainly

provides confidentiality to secret information, but can also provide authentication and nonrepudiation in some cases [12]. Steganography can also provide confidentiality of information, as well as privacy [13]. Steganography also provides security through obscurity, not always a good thing, but can be seen as a positive aspect in this case, since it is not the only means of security [14]. Importantly the proposed system provides a way of combining the two disciplines without increasing the amount of overhead used from the amount of overhead that a single encryption, or steganography, transaction would require.

A rather debatable advantage is that the proposed system makes provision for the use of proprietary encryption algorithms. Proprietary encryption algorithms are in most cases considered to be weak [15], since many get compromised due to inefficient algorithms. This aspect set aside, there are still many companies and individuals that prefer to use their own proprietary encryption algorithms to standard encryption algorithms. In the proposed system the fact that the algorithm is hidden inside the image increases the security of the algorithm and makes the distribution of the decryptor software more secure.

Another advantage of the proposed system is that it applies the diversity of defense concept [14], since it makes use of various layers of security. A lower security level steganographic algorithm can be used to embed the program and a higher security level steganographic algorithms can be used to embed the message.

4.2 Potential Weaknesses of the Proposed System

The first and most obvious risk to the proposed system is the fact that the decryptor and the encrypted message are stored in close proximity to one another. There are two possible solutions to this potential problem: Firstly one can divide the decryptor and the message between two different images. Embed the decryptor program in one image and embed the encrypted message in another and communicate them separately. As long as the decryptor program is capable of extracting and decrypting the message from the separate image file, the system will still function correctly. The second solution is to make use of cryptographic keys in the encryption of the message. Should someone try to execute the decryptor he will still need the secret key. Both of these solutions however will create more overhead, since more information needs to be communicated beforehand.

Another potential weakness lies in the way that the filename and file size are stored. Should an executable program be used for reasons of randomness, the filename and file size will still need to be in plaintext. This could provide valuable insight to an attacker who is trying to figure out what the true purpose of the hidden information is. A possible solution to the problem is to first encrypt at least the filename with a different encryption algorithm before it is embedded into the image. This approach however will create more overhead since the receiver must now again have knowledge of the encryption algorithm used in order to decrypt the filename.

Ultimately there exists a trade-off between the amount of overhead involved and the amount of security. More security could mean more unnecessary overhead, while less overhead will result in less security. It will ultimately depend on the desired level of security.

5 Conclusion

In trying to overcome the limitations that both encryption and steganography entail, a system is proposed that combines the two technologies in such a way as to minimise the amount of overhead used. This is done by embedding the decryptor program in the image along with the encrypted message.

The advantages that this approach offer include confidentiality and privacy of not only the secret message, but also potentially of the encryption algorithm. This results in other benefits that can be obtained, for example the secure use of proprietary encryption algorithms.

There are potential weaknesses to the system – most of their solutions include more overhead – and this brings about a trade-off between overhead and security. Ultimately, in whatever way the problems are dealt with, the proposed system will still involve less overhead than any similar security level combination of encryption and steganography.

References

- [1] Dunbar, B., “Steganographic techniques and their use in an Open-Systems environment”, *SANS Institute*, January 2002
- [2] Jamil, T., “Steganography: The art of hiding information is plain sight”, *IEEE Potentials*, 18:01, 1999
- [3] Silman, J., “Steganography and Steganalysis: An Overview”, *SANS Institute*, 2001
- [4] Johnson, N.F. & Jajodia, S., “Steganalysis of Images Created Using Current Steganography Software”, *Proceedings of the 2nd Information Hiding Workshop*, April 1998
- [5] Wang, H & Wang, S., “Cyber warfare: Steganography vs. Steganalysis”, *Communications of the ACM*, 47:10, October 2004
- [6] Johnson, N.F. & Jajodia, S., “Steganalysis: The Investigation of Hidden Information”, *Proceedings of the IEEE Information Technology Conference*, 1998
- [7] Krenn, R., “Steganography and Steganalysis”, <http://www.krenn.nl/univ/cry/steg/article.pdf>
- [8] Morkel, T., Eloff, J.H.P. & Olivier, M.S., “An overview of Image Steganography”, *Proceedings of the Information Security South Africa (ISSA) Conference*, 2005
- [9] Rubin, A.D. & Geer, D.E., “Mobile Code Security”, *IEEE Internet Journal*, December 1998
- [10] Linn, C. & Debray, S., “Obfuscation of Executable Code to Improve Resistance to Static Disassembly”, *Proceedings of the 10th ACM Conference on Computer and Communications Security*, 2003
- [11] “Obfuscated code”, *Wikipedia online encyclopedia*, http://www.wikipedia.org/wiki/Obfuscated_code, accessed on 6 July 2007
- [12] Tudor, J.K., “Information Security Architecture: An Integrated Approach to Security in the Organization”, *Auerbach Publications*, 2001, *book*
- [13] Artz, D., “Digital Steganography: Hiding Data within Data”, *IEEE Internet Computing Journal*, June 2001
- [14] Conklin, A., White, G.B., Cothren, C., Williams, D. & Davis, R.L., “Principles of Computer Security: Security+ and Beyond”, *McGraw-Hill Technology Education*, 2004, *book*
- [15] Schneier, B., “Security in the Real World: How to Evaluate Security Technology”, *Computer Security Journal*, Number 4, 1999

An Efficient Dispute Resolving Method for Digital Images*

Yunho Lee, Heasuk Jo, Seungjoo Kim**, and Dongho Won

Information Security Group,
Sungkyunkwan University,
300 Chunchun-dong, Suwon, Gyeonggi-do, 440-746, Korea
{leeyh, hsjo, skim, dhwon}@security.re.kr
<http://www.security.re.kr>

Abstract. Resolving rightful ownerships of digital images is an active research area of watermarking. Though a watermark is used to prove the owner's ownership, an attacker can invalidate it by creating his fake original image and its corresponding watermark. This kind of attack is called ambiguity attack and can be tackled either by use of non-invertible watermarking schemes or by use of zero-knowledge watermark detections. If a non-invertible watermarking scheme is used, then the owner should reveal her original image which should be kept secret. And if a zero-knowledge watermark detection is used, then no one can verify the claimed ownership unless the owner is involved. Moreover, in case of zero-knowledge watermark detection, the protocol is relatively complicated and needs more computations. In this paper, using the MSBs string of the original image other than the original image itself, we propose an efficient dispute resolving method while preserving secrecy of the original image.

Keywords: Ownership, Non-invertible Watermark, Similarity Functions.

1 Introduction

A watermark is inserted into digital images to prove the owner's copyright. For this reason, a watermark should not be removable or modifiable unless the perceptual quality of a digital image degraded severely and should be detected even after various signal processing such as JPEG compression or median filtering. This property of watermarking scheme is called *robustness*.

However, it is well known that a proof of presence of a certain watermark in a certain digital image can not be used to prove one's ownership of it. Because anyone can embed his watermark into a digital image that is already embedded with other watermark without sacrificing perceptual quality. Obviously the only way to resolve rightful ownership of a digital image containing multiple watermarks is to verify the disputants original images. Suppose that the true owner

* This work was supported by the University IT Research Center Project funded by the Korea Ministry of Information and Communication.

** Corresponding author.

A and an attacker B claimed ownership of the digital image \hat{I} . In this case, though the attacker's watermark W' can be detected in \hat{I} , there is no chance of detection of W' in A 's original image I . In opposite case, A 's watermark W should be detected in the attacker's claimed original image I' as it is derived from \hat{I} .

However, if an attacker can make his fake watermark that is detected in I , then he can invalidate the A 's ownership of that image. S.Craver *et al.* addressed this important issue first[1,4] and showed that the counterfeit watermarking schemes which can be performed on a watermarked image to allow multiple claims of ownership. Their attack method is known as *protocol attack* or *ambiguity attack*. To tackle this problem, they proposed a watermarking scheme which is claimed to be *non-invertible* and showed an instance of non-invertible watermarking scheme using the popular watermarking method proposed by I.J.Cox *et al.*[3]. They suggested that the watermark should be generated from the original image using one-way functions such as hash functions or symmetric ciphers. However, M.Ramkumar and A.N.Akansu broke their scheme more efficiently than a naive brute-force attack[6]. Please note that their successful attack does not mean that they broke the the cryptographic functions.

Subsequent research mainly focused on countermeasures against the attack proposed by S.Craver *et al.* and tried to apply cryptographic techniques such as hash functions, digital signatures and symmetric ciphers for dispute resolving [2,5]. On the other hand, there are a number of works exploiting weaknesses of claimed non-invertible schemes[6,9,11]. [9,11] give a formal definition of ambiguity attacks and argue that most proposed non-invertible schemes either do not come with a satisfactory proof of security, or the proofs are flawed.

Due to the difficulty of obtaining a non-invertible watermarking scheme, A.Adelsbach *et al.* proposed to use a trusted third party in order to generate watermark[9]. They claimed that non-invertible watermarking schemes only based on the one-way functions without a trusted third party are not provably secure because of the probability of "false alarms" of detecting functions. However, in 2004, Q.Li and E.C.Chang proposed a provably secure non-invertible watermarking scheme using a cryptographically secure pseudo-random number generator(CSPRNG) and proved that if their scheme is invertible, then the CSPRNG is insecure[13].

Non-invertible watermarking schemes require the true owner to give all information necessary for detection, for e.g., the watermark, the watermark detection key and even the original image to a judge or a dispute resolver. This requires an unnecessary high level of trust in a judge and we argue that the original image should be kept secret at least. The main reason why the original image I is required for dispute resolving is to prove its similarity to the disputed work \hat{I} , $I \approx \hat{I}$. Please note that blind watermarking schemes are not as suitable for proving similarity between images as non-blind watermarking schemes[10]. The same holds for asymmetric watermarking schemes, since they are inherently blind.

There is another approach in the context of dispute resolving which is based on the well known cryptographic technique, zero-knowledge proof[7,8,10,11,12].

[10] involves zero-knowledge watermark detection and timestamping authorities. They argue that they advocate the use of zero-knowledge watermark detection and time-stamping service for dispute resolving, since their security is better analyzed. Though zero-knowledge based approach is more secure than non-invertible watermarking schemes, no one can verify ownership of a digital image without involving the true owner. Obviously it will be more efficient that anyone can be easily convinced the claimed ownership is valid by her own computations without communicate with the owner.

Thus, the requirements of dispute resolving can be listed as follows.

1. *Secrecy*. The original image should be kept secret even to a judge or a dispute resolver.
2. *Public Verifiability*. Anyone can verify the claimed ownership of a certain digital image without involving the true owner.

Our Contribution. The second requirement can not be satisfied if we use zero-knowledge watermark detection scheme. Dispute resolving can be done by the following sequences[10].

1. Each disputant proves his creation of an original image.
2. Each disputant proves that her original image is similar to the disputed image.
3. Each disputant proves his creation time of the original image.

For the phase 1 and 2, we propose to use MSBs(Most Significant Bits) string of I and \hat{I} rather than I and \hat{I} for testing similarity. We believe that the secrecy of I can be preserved even if its MSBs string $MSB(I)$ is revealed, and that $MSB(I) \approx MSB(\hat{I}) \Leftrightarrow I \approx \hat{I}$ holds.

In this case, it is highly unlikely that the $MSB(I)$ is exactly the same as the $MSB(\hat{I})$ because of the watermark embedding induced distortion. Thus we propose more accurate similarity testing function than the previous ones.

For the phase 3, we incorporate timestamping authority as [10] to produce a watermark. Note that our dispute resolving method is independent to the underlying watermarking scheme and we assume that the underlying watermarking scheme is robust and blind one, for e.g., [14].

The rest of this paper is organized as follows. Section 2 introduces zero-knowledge watermark detection and non-invertible watermarking scheme, Section 3 proposes dispute resolving method using MSBs which is used to generate the watermark, and Section 4 presents security and performance analysis. Finally, this paper is concluded in Section 5.

2 Related Works

2.1 Zero-Knowledge Watermark Detection

Zero-knowledge proof is an interactive method for one party to prove to another that a statement is true, without revealing anything other than the veracity of the

statement. In 2000, S. Craver[7] proposed a secure watermark detection method using zero-knowledge protocol. In 2001, A. Adelsbach *et al.* proposed zero-knowledge watermark detection and proof of ownership method. Zero-knowledge watermark detection enables a prover(owner) to prove to an untrusted verifier that a certain watermark is present in stego-data without revealing any sensitive information for e.g., the watermark, the detection key, and the most sensitive information, the original image. Combining with time-stamping service, zero-knowledge watermark detection is also used for dispute resolving and authorship proof[10].

Though zero-knowledge watermark detection is useful and provably secure for dispute resolving and authorship proof, no one can verify the claimed ownership of a certain digital image without involving the true owner. Due to its lack of public verifiability, non-invertible watermarking scheme is considered more practical for dispute resolving if we can keep the original image's secrecy.

2.2 Non-invertible Watermarking Scheme

Let $\mathbb{W} = (\mathcal{E}, \mathcal{D}, \text{sim})$ be a watermarking scheme where $\mathcal{E}(\cdot)$ is an embedding function, $\mathcal{D}(\cdot)$ is a detecting or extracting function, and $\text{sim}(\cdot, \cdot)$ is a similarity testing function. Denote the A 's original image by I , the A 's watermark by W and the watermarked image by \hat{I} . A watermarking scheme \mathbb{W} should satisfying the following conditions,

$$\mathcal{E}(I, W, k_{e_A}) = \hat{I}, \quad (1)$$

$$\mathcal{D}(\hat{I}, k_{d_A}) = \hat{W}, \text{ and} \quad (2)$$

$$\text{sim}(\mathcal{D}(\hat{I}, k_{d_A}), W) = \text{TRUE} \quad (3)$$

where k_{e_A} and k_{d_A} are watermark embedding and detecting keys respectively.

Assume that B wants to invalidate the ownership of the image using invertible property of the watermarking scheme. Given \hat{I} , if B can successfully and efficiently find a watermark \tilde{W} and a fake original \tilde{I} such that

$$\text{sim}(\mathcal{D}(\hat{I}, \tilde{k}), \tilde{W}) = \text{TRUE} \wedge \text{sim}(\mathcal{D}(\tilde{I}, \tilde{k}), \tilde{W}) = \text{TRUE} \quad (4)$$

for an arbitrary key \tilde{k} then, we say that the watermarking scheme is *invertible*, otherwise *non-invertible*. Invertible watermarking schemes are susceptible to protocol attacks or ambiguity attacks because an attacker B can make his fake original image \tilde{I} and watermark \tilde{W} . Several authors proposed watermarking schemes and claimed that their schemes are non-invertible. Most of them follow the general design principle depicted in Fig. 1 as Craver *et al.* presented in their paper[4].

In Fig. 1, secure one-way or trapdoor functions such as hash functions or symmetric ciphers can be used as watermark generator. Non-invertible watermarking schemes' security is based on the one-wayness of the watermark generator in such a way that an attacker can not make his fake original from the watermarked image \hat{I} .

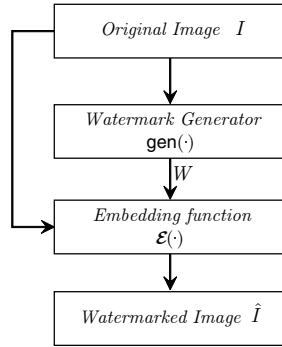


Fig. 1. General Design Principle for a Non-Invertible Watermarking Scheme

In general, the computation of fake original is believed to be hard because of the one-wayness of hash functions and symmetric ciphers. However, the security of these constructions is not based on the hash functions or symmetric ciphers but on the underlying watermarking scheme. Since watermarking schemes depend on statistical methods, the detecting function has a small probability of “false alarms.” [9] The “false alarms” means the detecting function indicates a watermark is believed to be embedded even though it was never embedded into the image. Thanks to “false alarms”, an attacker B choose a random \tilde{I} , computes the watermark $\tilde{W} = \text{gen}(\tilde{I}, \cdot)$ and checks whether \tilde{W} is detectable in \hat{I} . If this is not the case, he chooses another \tilde{I} [9].

This kind of attack is based on that the attacker can use the watermark generation function $\text{gen}(\cdot)$ in an unlimited manner. To avoid this problem, A.Adelsbach *et al.* proposed to construct the watermark in such a way that the computation of one single valid mark is already hard by incorporating digital signatures of a TTP(Trusted Third Party)[9].

3 Proposed Method

A non-invertible watermarking scheme requires both disputants to reveal their original images in order to resolve rightful ownership. Even if the original images are revealed to only a judge or a resolver, it requires a unnecessary high level of trust in them[10]. To remove this unnecessary high level of trust, we incorporate slightly different procedure for producing watermark that is to be embedded(See Fig. 2).

Note that the watermark is generated from the $\text{MSB}(I)$ rather than I itself where $\text{MSB}(\cdot)$ is a function of generating a bit string of MSBs in every bytes of image I . In this case, if the ownership of \hat{I} is challenged by attacker B , the true owner A reveals only the $\text{MSB}(I)$. A judge or a resolver verify that the following two conditions are true.

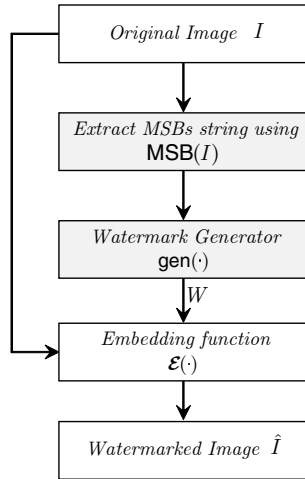


Fig. 2. Proposed Generation Process of Watermark using MSBs

- $MSB(I)$ is similar to $MSB(\hat{I})$ by similarity function $\text{sim}(\cdot, \cdot)$.
- the watermark \hat{W} extracted or detected from \hat{I} is derived from the $MSB(I)$.

These two checks ensure that the original image I is really exists that is similar to the disputed image \hat{I} . Obviously, as the $\text{sim}(\cdot, \cdot)$ plays an important role in deciding the similarity between $MSB(I)$ and $MSB(\hat{I})$, the $\text{sim}(\cdot, \cdot)$ needs to be accurate as possible. We will discuss how to make similarity function more accurate especially for comparison between two MSBs strings later.

Even if these two conditions are hold, it is insufficient for dispute resolving since those can not be used for proving the creation time of I . Thus, we use timestamping authority when producing watermark in order to prove the creation time of I .

3.1 The Watermark Generator

In addition to use $MSB(I)$ instead of I , we need a timestamping authority for proving creation time of I when producing a watermark. The watermark generation process of the owner A for original image I can be described in detail as follows(See Fig. 3).

1. A constructs $m = MSB(I)$ and sends it to the authority.
2. The authority computes $h_A = \text{hash}(ID_A, m)$, where $\text{hash}(\cdot)$ is a cryptographic one-way hash function and ID_A is A 's ID.
3. The authority produces a digital signature $W = \text{sign}(h_A, t_I)$ using her private key where t_I is a timestamp and $\text{sign}(\cdot)$ is a cryptographic message recovery digital signature method such as RSA and sends W to A .
4. A verifies whether $\text{verify}(W) = \text{hash}(ID_A || m) || t_I$ using the authority's public key. If it holds, A uses W as a watermark for I .

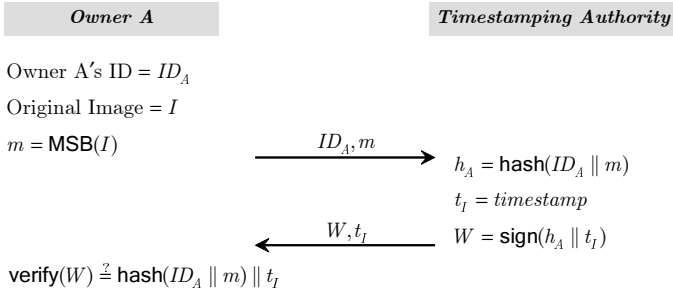


Fig. 3. The watermark generation process

3.2 Ownership Resolving Process

Suppose that there is one watermarked image \hat{I} and two watermarks W and \tilde{W} of both A and B respectively have been extracted(detected) and both of them are claiming to be its rightful owner. In this case, a judge require both of them to reveal their MSBs strings of the original images. Let $m = \text{MSB}(I)$ be the A 's MSBs string and $\tilde{m} = \text{MSB}(\tilde{I})$ be the B 's MSBs string. For dispute resolving, a judge should check the followings.

1. A judge verifies two detected watermarks W and \tilde{W} using the authority's public key and produces MSBs strings and timestamps. If verification of W fails then B is the true owner and vice versa.

$$m' \parallel t_I \leftarrow \text{verify}(W) \tag{5}$$

$$\tilde{m}' \parallel t_{\tilde{I}} \leftarrow \text{verify}(\tilde{W}) \tag{6}$$

The judge also checks whether $m = m'$ and $\tilde{m} = \tilde{m}'$. If $m = m'$ and $\tilde{m} \neq \tilde{m}'$ then A is the true owner and vice versa.

2. If both verifications are succeeded then the judge calculates two similarity values δ_A and δ_B by

$$\delta_A = \text{sim}(\text{MSB}(\hat{I}), m) \tag{7}$$

$$\delta_B = \text{sim}(\text{MSB}(\hat{I}), \tilde{m}). \tag{8}$$

If only one value exceeds the predefined threshold, then the owner of the value is the true owner of \hat{I} .

3. If both of them are exceed the threshold, then both timestamp t_I and $t_{\tilde{I}}$ are compared. If $t_I < t_{\tilde{I}}$ then A is the true owner and vice versa.

For the check 1, we assume that the watermark detection is accurate enough and the watermark detection is wholly up to the underlying watermarking scheme and that is beyond the scope of this paper. We suggest that the underlying watermarking scheme should be blind and robust one.

3.3 A New Similarity Function for Comparing Two MSBs Strings

In general, the similarity of two vectors X and Y can be computed by (9)[3].

$$\text{sim}(X, Y) = \frac{X \cdot Y}{\sqrt{Y \cdot Y}}. \tag{9}$$

Though the similarity function $\text{sim}(\cdot, \cdot)$ can be used for testing similarity between two MSBs strings, we can make the function more accurate considering watermark embedding induced distortion.

An image can be represented by byte string. Let x be the byte string of I and y be the byte string of \hat{I} . And suppose that $m = \text{MSB}(I)$, $\hat{m} = \text{MSB}(\hat{I})$, m_i is the i -th bit of m , and \hat{m}_i is the i -th bit of \hat{m} . Due to the imperceptibility property of the underlying robust watermarking scheme, y_i , the i -th byte of y , is slightly different from x_i . This means that all the probabilities $\Pr[\text{MSB}(x_i) = \text{MSB}(y_i)]$ are not the same to each other for $i = 1, 2, \dots, |x|$.

Intuitively, if $x_i = 2^7$ or $x_i = 2^7 - 1$ then the probability will be the maximum and if $x_i = 0$ or $x_i = 255$ then the probability will be the minimum where $|x_i| = 8$. Our experiments show that about 0.6%–1.6% bits of \hat{m} are different from m (See Table 1.) in case of general QIM watermarking scheme with step size 9–45. Table 1. shows that the $\Pr[m_i \neq \hat{m}_i]$ is negligible if $y_i < 120$ and $135 < y_i$.

Table 1. The changing rate of the MSBs by watermark embedding induced distortion

Images(512×512)	Total MSBs	Changed MSBs	Rates	Value of $y_i(m_i \neq \hat{m}_i)$
Lena	262,144	1,697	0.65%	120–134
Baboon	262,144	4,133	1.58%	121–134
Peppers	262,144	3,467	1.32%	121–135

Thus it is more accurate that we should apply different weight according to y_i for computing similarity between m and \hat{m} .

We propose a new similarity function $\text{sim}^*(\cdot, \cdot)$ for computing similarity between two MSBs strings m and \hat{m} as (10).

$$\text{sim}^*(m, \hat{m}) = \sum_{i=1}^{|m|} (m_i \oplus \hat{m}_i) \frac{|y_i - 2^{k-1}|}{2^{k-1}}, (m_i, \hat{m}_i \in \{0, 1\}) \tag{10}$$

where \oplus means xor operation and k is the bit length of y_i . If $m = \hat{m}$ then $\text{sim}(m, \hat{m}) = 0$. In case of 512 by 512 image, if $m = \hat{m}$ and $m_i, \hat{m}_i \in \{-1, 1\}$ ($1 \leq i \leq 512^2$) then

$$\text{sim}(m, \hat{m}) = \frac{m \cdot m}{\sqrt{m \cdot m}} = 512, \text{ and} \tag{11}$$

$$\text{sim}^*(m, \hat{m}) = \sum_{i=1}^{|m|} (\hat{m}_i \oplus m_i) \frac{|y_i - 2^{k-1}|}{2^{k-1}} = 0. \tag{12}$$

4 Security and Performance Analysis

4.1 Security Analysis

Though an attacker can not have the original image I , he can make a fake original image \hat{I} which is similar to the disputed image \hat{I} . However, the attacker can not get timestamp from the authority prior to the true owner unless he can predict $MSB(I)$ before the true owner create it.

The only way to invalidate ownership is to make possible bit strings and get timestamps from the authority. Though this looks like well-known brute force attack, assume that an attacker has a polynomial function $f(x)$ where x is a bit string which can distinguish between random bit strings and MSBs strings from digital images, then the attack will be more efficient than the naive brute force.

However, we believe that MSBs strings from digital images and random bit strings are polynomially indistinguishable though we can not prove formally yet.

4.2 Performance of the New Similarity Function

The accuracy of two similarity functions $\text{sim}(m, \hat{m})$ and $\text{sim}^*(m, \hat{m})$ is depicted in Table 2 in case of the dimension of \hat{I} is 512 by 512. If there is no transmission error such as JPEG compression, the accuracy of $\text{sim}(\cdot, \cdot)$ is about 97.6% and the accuracy of $\text{sim}^*(\cdot, \cdot)$ is 99.9%.

Table 2. The Results of the two similarity functions $\text{sim}(\cdot, \cdot)$ and $\text{sim}^*(\cdot, \cdot)$

Case	$\text{sim}(\cdot, \cdot)$	$\text{sim}^*(\cdot, \cdot)$
$m = \hat{m}$	512	0
$m \approx \hat{m}$	495–505	6–14
$m \neq \hat{m}^1)$	≈ 0	$\approx 30,000$
Accuracy in case of $m \approx \hat{m}$	97.6%	99.9%

1) m is created independently from \hat{m} .

Table 3. The results of $\text{sim}(\cdot, \cdot)$ and $\text{sim}^*(\cdot, \cdot)$ with 50% JPEG compressed \hat{I}

Image	$\text{sim}(\cdot, \cdot)$	$\text{sim}^*(\cdot, \cdot)$
Lena	484.5(94.64%)	69.0(99.77%)
Baboon	441.4(86.21%)	436.4(98.55%)
Peppers	493.9(96.46%)	58.6(99.81%)

Table 3. shows the similarity results in case of JPEG 50% compressed \hat{I} . The results of $\text{sim}^*(\cdot, \cdot)$ is far better than the normal $\text{sim}(\cdot, \cdot)$ especially in Baboon image.

Fig. 4. shows the responses of the similarity functions $\text{sim}(\cdot, \cdot)$ and $\text{sim}^*(\cdot, \cdot)$ to 1,000 randomly generated MSBs of which only one matches \hat{m} respectively.

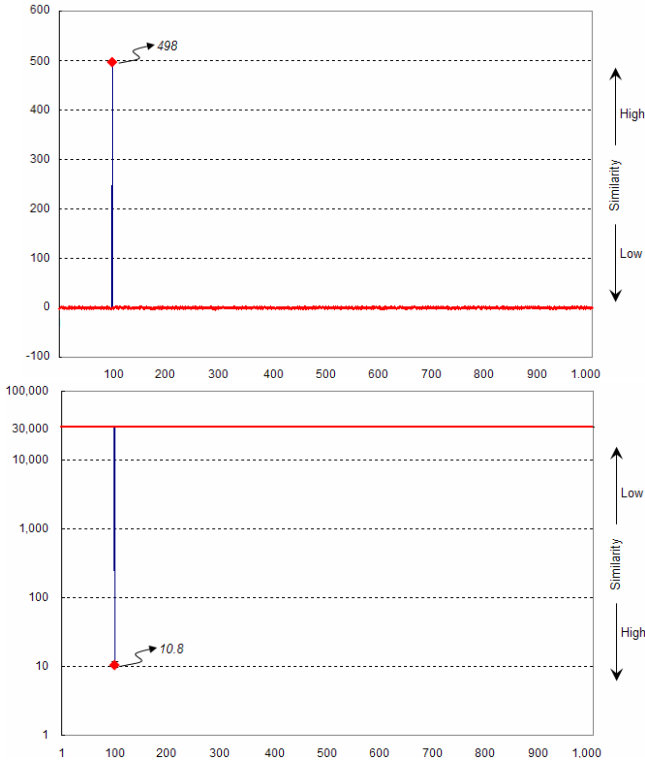


Fig. 4. The similarity results of $\text{sim}(\cdot, \cdot)$ (top) and $\text{sim}^*(\cdot, \cdot)$ (bottom)

5 Concluding Remarks

A non-invertible watermarking scheme is based on the one-way function which is used for generating watermark from the original image. However, in order to prove ownership the true owner should reveal the original image because the embedded watermark is generated from the original image. In this paper, we propose to use the MSBs of the original image I rather than the original image I itself and a more accurate similarity function for comparing $\text{MSB}(I)$ and $\text{MSB}(\hat{I})$ than the ordinary similarity function.

References

1. S.Craver, N.Memon, B.Yeo, and M.Yeung, "Can Invisible Watermarks Resolve Rightful Ownership?," *IBM Research Division, Tech. Rep.*, RC20509, 1996.
2. W.Zeng and B.Liu, "On Resolving Rightful Ownerships of Digital Images by Invisible Watermarks," *Proc. of the 4th International Conference on Image Processing(ICIP)*, pages 552–555, 1997.

3. I.J.Cox, J.Kilian, F.T.Leighton, and T.Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," *IEEE Trans. on Image Processing*, vol.6, pages 1673–1687, 1997.
4. S.Craver, N.Memon, B.Yeo, and M.Yeung, "Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications," *IEEE J. on Selected Areas of Communications*, vol.16, no.4, pages 573–586, 1998.
5. L.T.Qiao and K.Nahrstedt, "Watermarking Schemes and Protocols for Protecting Rightful Ownership and Customer's Rights," *Journal of Visual Communication and Image Representation*, vol.9, no.3, pages 194–210, 1998.
6. M.Ramkumar and A.N.Akansu, "Image Watermarks and Counterfeit Attacks: Some Problems and Solutions," *Proc. of Content Security and Data Hiding in Digital Media*, May 1999.
7. S.Craver, "Zero Knowledge Watermark Detection," *Proc. of Information Hiding 99(IH99)*, LNCS 1768, 2000.
8. A.Adelsbach and A.R.Sadeghi, "Zero-Knowledge Watermark Detection and Proof of Ownership," *Proc. of the 4th International Workshop on Information Hiding*, LNCS 2137, pages 273–288, 2001.
9. A.Adelsbach, S.Katzenbeisser, and H.Veith, "Watermarking Schemes Provably Secure Against Copy and Ambiguity Attacks," *Proc. of the 2003 ACM Workshop on DRM*, pages 111–119, 2003.
10. A.Adelsbach and A.R.Sadeghi, "Advanced Techniques for Dispute Resolving and Authorship Proofs on Digital Works," *Proc. of SPIE Security and Watermarking of Multimedia Contents V*, vol.5020, 2003.
11. A.Adelsbach, S.Katzenbeisser, and A.R.Sadeghi, "On the insecurity of non-invertible watermarking schemes for dispute resolving," *Proc. of International Workshop on Digital Watermarking 2003(IWDW2003)*, LNCS 2939, pages 355–369, 2003.
12. A.Adelsbach, S.Katzenbeisser, and A.R.Sadeghi, "Watermark Detection with Zero-Knowledge Disclosure," *ACM Multimedia Systems Journal, Special Issue on Multimedia Security*, Vol.9, No.3, pages 266–278, Sep. 2003.
13. Q.Li and E.C.Chang, "On the Possibility of Non-Invertible Watermarking Schemes," *6th International Workshop on Information Hiding*, LNCS 3200, Springer-Verlag, pages 13–24, 2004.
14. Y.Lee, K.Lee, S.Kim, D.Won, and H.Yang, "A Method for Deciding Quantization Steps in QIM Watermarking Schemes," *The 1st International Workshop on Security in Ubiquitous Computing Systems(in conjunction with EUC 2005)*, Springer-Verlag, LNCS 3823, pages 965–975, 2005.

Use of SAML for Single Sign-On Access to Multimedia Contents in a Peer-to-Peer Network

Rubén Barrio, Xavier Perramon, and Jaime Delgado

Universitat Pompeu Fabra

Pg. Circumval·lació 8, E-08003 Barcelona, Spain

{ruben.barrio, xavier.perramon, jaime.delgado}@upf.edu

Abstract. The Single Sign-on technique (SSO) facilitates the management of authentication and authorization functions in a federation of servers providing access to protected resources. Different approaches to SSO exist, among them the XML-based Security Assertion Markup Language (SAML) standard, which has been used in applications such as intranet access within organizational domains. This paper focuses on the use of SAML for authentication and authorization in a project aimed at providing peer-to-peer access to high-definition audiovisual streams. Adaptation of various elements are proposed herein in order to cope with the specific characteristics of SSO in a P2P architecture.

1 Introduction

Users of a federation of service providers will usually need to authenticate themselves before accessing any resource afforded by these providers. The goal of the Single Sign-on technique (SSO) is that, within an identity federation, the process of authentication needs to be carried out only once during an access session, no matter how many service providers take part in this session. Typical usage scenarios include a user booking flight tickets, hotel rooms and car rentals in a single web access session, even though several servers may be involved, or students and university staff accessing various servers in an online campus through a single log-in.

A new application scenario where SSO may be required is the sharing of resources in a peer-to-peer network, where access to such resources is restricted to users having the appropriate authorization. Although P2P networks have traditionally been used for totally unrestricted file sharing, sometimes even violating property rights, their architecture is also suitable for applications with stricter access control requirements. Thus, users accessing restricted resources, e.g. files they have paid for, can also benefit from the P2P paradigm.

In this paper we present the MACHINE project, whose main goal is to build peer-to-peer networks for the interchange of high-definition multimedia files. One of the requirements for this network is that access to these files must be restricted to authorized users, but they need authenticate only once to the system, i.e. a SSO functionality is required. Next, we present the basis for the SSO approach implemented in the MACHINE project, namely the SAML standard [1]. In the

following sections we describe how SSO can be applied to a P2P environment, and in particular how we are using SAML in the MACHINE project. Finally, some security aspects are considered based on previous security analyses of the SAML standard, and conclusions are drawn from this implementation of SSO in a P2P network.

2 The MACHINE Project

2.1 Overview

MACHINE is a project funded by i2CAT [2], a Foundation comprising the Catalan Government and several universities, among them UPC (Universitat Politècnica de Catalunya) and UPF (Universitat Pompeu Fabra). The main goal of the project is to facilitate the interchange of high-definition multimedia contents based on a peer-to-peer paradigm (P2P). The MACHINE system includes a resource discovery service that allows any peer to locate a specific multimedia file. Once this resource is located, and if the requesting user is properly authorized, a multimedia stream is transmitted for real-time visualization.

As is customary in P2P systems, the desired resource may be located in more than one peer. One of the features of MACHINE is that not only the whole multimedia files can be replicated, but also fragments of these files. This allows for sequential download from different peers, according to the network conditions, but also parallel download, e.g. different layers of a video stream and/or the corresponding audio stream can be transmitted from different sources simultaneously.

The MACHINE project is the successor of the “i2CAT Projecte Integrat” (I2CatPI). The I2CatPI consisted of a prototype peer-to-peer system for multimedia distribution, and one of its requirements was also SSO, but in that project version 1 of SAML was used as the standard for representing authentication information. A simple profile was used based on user name and password for authentication with an Authentication Entity, which returned a signed assertion with the name of the user and the corresponding system role. Afterwards, the different service providers of the system could verify the authenticity of this assertion by means of the signature issued by the authentication server. In the I2CatPI project we implemented a simple SSO system with a basic security level, and in the new MACHINE project we are extending the functionalities of this predecessor.

2.2 Security Requirements

Resources stored in this P2P network will generally have access restrictions, i.e. not everyone will be allowed to access any multimedia stream. Therefore, an authentication and authorization mechanism needs to be used in the system. Since there may be more than one provider peer involved in the transfer, various authentication processes might be necessary during a session. It is a requirement of the system that this can be done transparently to the user, using an

SSO approach, so that authentication needs to be carried out only once, and the identification and authorization information is conveniently transferred to whichever nodes may require it.

It is also a requirement that the SSO method used must be as secure as possible in order to prevent unauthorized accesses. Therefore, not only a secure authentication scheme must be implemented, but also the transfer of the streams must be protected with encryption techniques, adapted to the specific characteristics of efficient transmission of high-definition multimedia contents.

Another requirement is that the solution adopted should be based on standards, in order to facilitate interoperability. The standard chosen for fulfilling these requirements for authentication and authorization is the Security Assertion Markup Language (SAML).

3 The SAML Standard

The SAML standard consists of a series of specifications defining the format of messages, structured as XML elements, to be exchanged in authentication and authorization protocols. It is developed and maintained by the Security Services Technical Committee (SSTC) of the OASIS Consortium. Different versions of the standard have been published so far: version 1.0 was released in 2002, and a revision thereof, version 1.1, in 2003. The current version is 2.0 [1], published in 2005, which is not fully backwards compatible with the previous versions. Unless explicitly stated, all references to the SAML standard in this paper must be understood as referring to the latest version, i.e. version 2.

The initial goal of SAML was to provide a mechanism for Single Sign-on in an environment where various servers can provide services to an identified user. The most usual situation would be that of a user with a common web browser accessing web servers in federated identity domain. SAML provides mechanisms whereby the browser can automatically send authentication and authorization information between servers, with the user only needing to explicitly authenticate before one of these servers.

This usage of SAML for SSO has been deployed in a number of projects. One of the most known of these is the open source Shibboleth project for inter-organization sharing of web resources e.g. between educational institutions [3].

Various types of items are defined in the SAML specifications: *assertions*, *protocol messages*, *protocol bindings* and *profiles*.

3.1 SAML Assertions

The XML `<Assertion>` element is the basis of the SAML standard. It contains sub-elements such as `<Issuer>`, an optional `<Signature>`, and a sequence of *statements*. The assertion must be authenticated by its issuer, and this can be done with the `<Signature>` element, in accordance with the XML Signature specification [4], or by external means, e.g. within a communications protocol already providing authentication. If necessary, assertions or elements thereof can also be encrypted with the XML Encryption technique [5].

Most assertions contain a `<Subject>` element, identifying the subject of the statement(s) in the assertion. This is usually done through a sub-element of type `<SubjectConfirmationData>`, containing information to be used in the confirmation of the subject, such as an IP address, a public key, a certificate, etc.

The statements within an assertion can be of various types, the most remarkable being authentication statements (`<AuthnStatement>` element) and authorization statements (`<AuthzDecisionStatement>` element). This latter type is being retained in the standard for compatibility with previous versions, with the use of a more flexible alternative, as defined in the new XACML standard [6], being encouraged.

3.2 SAML Protocol Messages

SAML also defines two basic types of elements to be used as messages in an authentication and/or authorization protocol: requests and status responses. Requests can also be of various types, e.g. `<AuthnRequest>` for requesting authentication and `<AuthzDecisionQuery>` for requesting authorization. Successful responses will usually include assertions containing `<AuthnStatement>` and `<AuthzDecisionStatement>` elements respectively, as described above.

3.3 SAML Protocol Bindings and Profiles

SAML protocol messages and their enclosed assertions can be used in different application contexts. One of the specifications of the SAML standard defines the so-called *protocol bindings*, i.e. mappings of SAML protocol messages onto communication protocols. Examples of such bindings are the exchange of SAML messages within SOAP messages over HTTP, through HTTP Redirect messages, or through HTTP POST operations, all of these with or without an underlying SSL/TLS secure connection, depending on the application.

Another binding defines the exchange of SAML messages by reference, where the references to the SAML information are called *artifacts*. These are to be used in a context where the other bindings are inappropriate due to the length of the SAML message (e.g. in the HTTP Redirect binding the whole SAML message has to be encoded in the redirected URL) or to other reasons, such as the occurrence of an untrusted proxy. Such artifacts are then resolved through a separate binding, which can be e.g. SOAP over HTTP.

For the specific types of applications that may make use of SAML, the standard defines a number of *SAML profiles* [7], i.e. sets of rules for the exchange of SAML messages using a specific binding. The basic profile, called the Web Browser SSO profile in SAML v. 2, specifies how to achieve Single Sign-on in a federation of web servers with an unmodified web browser. The standard also defines other profiles for SSO and for artifact resolution, and also provides guidelines for the specification of new profiles, adapted to other scenarios.

4 Single Sign-On Access to P2P Resources

As explained in the previous section, the main goal of the initial versions of the SAML specification was to facilitate SSO access to different web servers with a normal web browser. Subsequently, SAML profiles have been defined for other scenarios, such as attribute-based authorization or secure web services.

In the context of the MACHINE project, the SSO feature is needed for accessing protected resources in a P2P environment. These resources consist of multimedia contents, which are possibly distributed among various peers in the network. This allows for better efficiency in the transfer of such contents: they can be fragmented and/or replicated in different hosts, so that the audiovisual information can be streamed from multiple peers sequentially or even in parallel.

Obviously, users will have to authenticate themselves before accessing a protected stream. The authentication and authorization process can be performed during the resource discovery phase, or directly at the start of the transmission if e.g. the user knows beforehand the location of the desired contents. In either case, the authorization information will be securely conveyed to all other peers involved in the transfer in a manner completely transparent to the user, as expected in a Single Sign-on environment.

The natural implementation of SSO in this P2P system is based on a central authorization server capable of authenticating and issuing authorization statements about registered users, and trusted by all peers belonging to the P2P domain. Once a session is established, the first time authorization is required by any provider (i.e., peer) the necessary steps are carried out between this provider, the central server, and the user through the corresponding user agent. Any other provider requiring authorization information during the same session will obtain it by interaction with the central server, without user intervention.

This scenario differs from the traditional SSO web access in the following aspects:

- Web SSO techniques are typically designed for use with a generic client, i.e. a plain web browser not necessarily aware of any SSO mechanisms. This is usually achieved by exchanging state information between servers, e.g. by means of HTTP cookies, HTTP Redirect messages, or automatic submission of forms through JavaScript. In our P2P case, the stream transfer system is specifically designed for coping with the distributed authorization scheme.
- Some web SSO techniques may be valid only within a domain, as it is the case for HTTP cookies. According to the cookies specification [8], any HTTP cookie is to be sent only to servers within a given DNS domain, thus precluding submission of state information to servers outside this domain. In our P2P scenario, there is no restriction on the DNS domains where the peers may be located.
- In web SSO, there is a clear role distinction between clients and servers, the latter being also distinguished between service providers and authentication servers. In our system, because of its peer-to-peer nature, every host can potentially be an authorization requester (as web service providers are) and/or an authorization presenter (as web clients are).

5 Use of SAML in the P2P Environment

In order to implement SSO authentication and authorization in the P2P system of the MACHINE project, several approaches may be considered. Given the availability of SSO mechanisms for web access, one possible approach would be based on the exchange of HTTP messages between the peers. Examples of such mechanisms are the abovementioned HTTP cookies, HTTP Redirect messages, and forms, and also the Web Browser SSO Profile of SAML [7]. All of these, however, are designed to be used with a standard web browser with no added functionality. If specific software modules or plug-ins are developed for the client side, as it is the case in the MACHINE project, it is preferable to employ them for better efficiency and flexibility.

The solution adopted in the MACHINE project is based on SAML, but without using the standard Web Browser SSO Profile. We can consider our solution as a new SAML profile, according to [7], specific to the P2P environment.

5.1 The MACHINE Scenario

Figure 1 depicts the elements that make up the MACHINE scenario. The MACHINE specification provides for the definition of *P2P domains*. Among other components, a P2P domain is formed by a *Resource Discovery Server* (RDS), an *Authentication and Authorization Server* (AAS), and a number of hosts acting as peers. The RDS and the AAS may be located on the same host. The local storage of each peer may contain multimedia files and/or fragments thereof, to be streamed to other peers upon request (after querying the RDS if necessary). There is no constraint on the location of the servers or the peers, neither regarding network topology, nor in particular regarding DNS domain names.

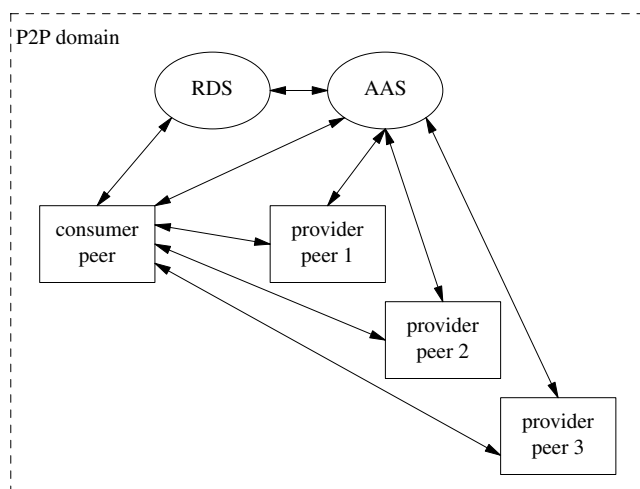


Fig. 1. The MACHINE P2P scenario

Every peer willing to join a P2P domain must register with the AAS for that domain. As part of the registration process, the peer must present a public key or a valid X.509 certificate. In the former case, the AAS itself may issue a certificate for the peer. In the latter case, the AAS will make use of a local database of acceptable Certification Authorities. These certificates will be used, on one hand, for the SAML Authentication Request Protocol as defined in [1] (see Sect. 5.2), and on the other hand, for establishing mutually authenticated SSL/TLS connections for the interchange of SAML messages, as discussed in Sect. 6.

The AAS also has a database containing an access control list (ACL) to resources. This is used when issuing authorization statements, i.e. the ACL is looked up to check whether a peer has permission to access a given multimedia stream. Alternatively, the access to multimedia contents may be governed by DRM licenses associated with them. One of the activities of the MACHINE project consists in mapping the rights expressed with standard techniques in such licenses to the authorization statements described here.

5.2 The P2P SAML Profile

When a peer, hereinafter called the *consumer peer*, wants to access a resource in the P2P domain, it must use the service provided by the RDS or otherwise obtain the locator for that resource. This includes information on the peer or set of peers capable of providing the desired resource, i.e. the *provider peer(s)*.

The RDS can submit a SAML authentication request to the AAS in order to verify the identity of the consumer peer, and subsequently a SAML assertion query in order to check whether this peer is authorized to access the resource. Or, alternatively, these steps can be carried out by the provider peer (this will be the case when the consumer peer does not make use of the RDS). Note that, in the first case, the RDS and the AAS could be located on the same host.

In the following description, the entity (either the RDS or the provider peer) requesting authentication and authorization information about the consumer peer is called the *requester*, according to the terminology used in [1].

When the requester determines that authentication and authorization must be performed on a consumer peer, the following steps are carried out:

1. The requester sends a SAML `<AuthnRequest>` message to the AAS requesting authentication of the consumer peer. This message includes a `<Subject>` element containing the consumer peer's X.509 certificate, and also its IP address.
2. The AAS validates the authentication information. If this is the first authentication request for the subject in the current session, validation is performed through direct interaction with the consumer peer. This interaction consists in signing a randomly generated challenge with the peer's private key (and thus possibly requiring action from the user, e.g. entering the passphrase under which the private key is protected). After the signature is successfully checked with the certificate, the AAS sends a SAML `<Response>` message to the requester containing an `<AuthnStatement>` element.

3. If the AAS response confirms the authenticity of the consumer peer, the requester sends a SAML `<AuthzDecisionQuery>` message to the AAS querying whether the consumer peer is authorized to access the desired resource.
4. The AAS looks up its ACL database and sends a SAML `<Response>` message to the requester containing an `<AuthzDecisionStatement>` element.
5. Based upon the response received from the AAS, the requester grants or denies the desired service to the consumer peer, identified by the `<Subject>` element in the `<AuthzDecisionStatement>` assertion. If the requester is the RDS, the resource locator is or is not returned to the consumer peer, and if the requester is the provider peer, the multimedia is or is not streamed to the consumer peer.

The same consumer peer may have to be authenticated by other entities during the same session. This will happen when authentication has been performed by the RDS and the consumer peer accesses the (first) provider peer, or when more than one provider peer generate the stream to be transmitted to the consumer peer. In such cases, the steps described above are repeated by the new requester, generally with the same result (although it might be possible that the consumer's rights to the desired resource were revoked after the session is established: this revocation would then take effect only when a new peer is accessed). However, the interactive validation of step 2 above is not performed again, but rather the AAS returns directly the response based on the previous authentication, thereby achieving the ultimate goal of Single Sign-on.

6 Security Considerations

The SAML standard is designed for providing authentication and authorization statements in a secure manner, but also for minimizing the processing overhead where possible. For this reason, version 1 of the standard does not require the use of confidentiality and/or authentication mechanisms when they are not considered necessary.

The security afforded by the use of SAML has been object of various studies. In particular, Groß [9] showed that when the Browser/Artifact profile of SAML v. 1 is used, several attacks are possible in certain circumstances if some messages that the profile does not require to protect are indeed not protected. This is the case of unauthenticated messages to the identity provider, encrypted but unauthenticated HTTP redirects to the service provider containing SAML artifacts, or unencrypted error messages (possibly provoked by the attacker) containing HTTP Referrer headers. The main conclusion of this analysis was that external protection measures, such as sending all HTTP messages over SSL/TLS [10,11] connections (i.e. making use of HTTPS rather than plain HTTP), should be taken when the SAML profile does not provide enough security.

The committee responsible for the SAML standard acknowledged [12] that these types of attacks were possible and that, although the security context where SAML messages are used is out of the scope of the SAML standard, the

use of SSL/TLS is recommended in SAML v. 1, and this recommendation has been strongly emphasized in SAML v. 2.

In our implementation of SAML-based SSO for the P2P environment in the MACHINE project, all exchanges of SAML messages must always be carried over SSL/TLS connections with mutual authentication, i.e. with both client-side and server-side authentication. For this reason, all peers belonging to a P2P domain must register with their X.509 certificate.

Appropriate use of SSL/TLS prevents, or at least makes reasonably unfeasible, the following types of attack:

- Forgery. An attacker could fabricate an `<AuthzDecisionStatement>` assertion granting access to some resource and send it to the authorization requester. Since the message will not come from an authenticated SSL/TLS connection with the AAS, the requester will refuse it.
- Reply attacks. An attacker could capture an `<AuthzDecisionStatement>` assertion sent from the AAS to the authorization requester and send it to this requester pretending to be the legitimate consumer. In order for this attack to be successful, the attacker would have to be able to replace the `Address` attribute, containing the consumer's IP address, included in the `<SubjectConfirmationData>` element of the `<AuthzDecisionStatement>` message, which will be encrypted with the legitimate SSL/TLS session keys, unknown to the attacker.
- Man-in-the-middle attacks. These attacks are completely deterred when mutual authentication based on X.509 certificates is used in an SSL/TLS connection.
- Information leakage. No protocol information can be obtained by an attacker, e.g. from error messages in the protocol, because all types of messages will be conveyed over the SSL/TLS connections.

Apart from these communication-oriented attacks, supplantation is directly inhibited by the SAML authentication protocol (step 2 in Sect. 5.2 above).

7 Conclusions

We have shown how version 2 of the SAML standard can be deployed in a peer-to-peer network for providing Single Sign-on access to multimedia resources among various hosts acting as peers. SAML provides a framework for the representation and interchange of authentication and authorization information, and also defines a number of profiles to be used for SSO in different contexts, typically in web access to a federation of servers. But SAML also provides guidelines for the definition of new profiles, and we have adapted the usage of SAML to the MACHINE project, where P2P access to a multimedia resource can be done with the collaboration of various peers, and also of a resource discovery server.

Therefore, we have implemented SSO access within this P2P network environment by direct exchange of SAML protocol messages, without using the standard

web browser profile. We have thus defined a new SAML profile for use in this P2P scenario.

We have also analyzed how the combination of SAML protocol messages together with a secure communication layer, like the one afforded by the SSL/TLS protocols, provides a strong measure against possible fraud.

The use of SAML in this case contributes to enhanced security and interoperability, since the peer systems can make use of any SAML-based implementation for building and exchanging authentication and authorization messages.

This has shown that SAML provides enough flexibility to adapt it to different application scenarios, of which SSO access to a P2P network is a new example.

References

1. Cantor, S., Kemp, J., Philpott, R., Maler, E.: Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. <http://www.oasis-open.org/committees/security/>, Doc. ID `saml-core-2.0-os` (2005)
2. i2CAT web site. <http://www.i2cat.org/>
3. Erdos, M., Cantor, S.: Shibboleth Architecture. <http://shibboleth.internet2.edu/docs/draft-internet2-shibboleth-arch-v05.pdf> (2002)
4. Eastlake, D., Reagle, J., Solo, D.: XML-Signature Syntax and Processing. <http://www.w3.org/TR/xmlsig-core/> (also published as RFC 3275) (2002)
5. Eastlake, D., Reagle, J.: XML Encryption Syntax and Processing. <http://www.w3.org/TR/xmlenc-core/> (2002)
6. Moses, T.: Extensible Access Control Markup Language (XACML) Version 2.0. <http://www.oasis-open.org/committees/xacml/>, Doc. ID `access_control-xacml-2.0-core-spec-os` (2005)
7. Hughes, J., Cantor, S., Hodges, J., Hirsch, F., Mishra, P., Philpott, R., Maler, E.: Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0. <http://www.oasis-open.org/committees/security/>, Doc. ID `saml-profiles-2.0-os` (2005)
8. Netscape: Persistent Client State — HTTP Cookies. http://wp.netscape.com/newsref/std/cookie_spec.html (1999)
9. Groß, T.: Security Analysis of the SAML Single Sign-on Browser/Artifact Profile. 19th Annual Computer Security Applications Conference, Las Vegas (2003)
10. Netscape: SSL 3.0 Specification. <http://wp.netscape.com/eng/ss13/> (1996)
11. Dierks, T., Allen, C.: The TLS Protocol Version 1.0. RFC 2246 (1999)
12. Linn, J., Mishra, P.: SSTC Response to “Security Analysis of the SAML Single Sign-on Browser/Artifact Profile”. <http://www.oasis-open.org/committees/security/>, Doc. ID `sstc-gross-sec-analysis-response-cd-01` (2005)

EMAP: An Efficient Mutual-Authentication Protocol for Low-Cost RFID Tags

Pedro Peris-Lopez, Julio Cesar Hernandez-Castro,
Juan M. Estevez-Tapiador, and Arturo Ribagorda

Computer Science Department, Carlos III University of Madrid
{pperis, jcesar, jestevez, arturo}@inf.uc3m.es

Abstract. RFID tags are devices of very limited computational capabilities, which only have 250-3K logic gates that can be devoted to security-related tasks. Many proposals have recently appeared, but all of them are based on RFID tags using classical cryptographic primitives such as PRNGs, hash functions, block ciphers, etc. We believe this assumption to be fairly unrealistic, as classical cryptographic constructions lie well beyond the computational reach of very low-cost RFID tags. A new approach is necessary to tackle this problem, so we propose an extremely efficient lightweight mutual-authentication protocol that offers an adequate security level for certain applications and can be implemented even in the most limited low-cost RFID tags, as it only needs around 150 gates.

Keywords: Ubiquitous Computing, RFID, Tag, Reader, Privacy, Tracking, Pseudonym, Mutual-authentication.

1 Introduction

Low-cost Radio Frequency Identification (RFID) tags affixed to consumer items as smart labels are emerging as one of the most pervasive computing technologies in history. This presents a number of advantages, but also opens a huge number of security problems that need to be addressed before their successful deployment. The most important security questions are privacy and tracking, but there are some others worth to mention, such as physical attacks, denial of service, etc.

The low cost demanded for RFID tags (0.05-0.1€) forces the lack of resources for performing true cryptographic operations. Typically, these systems can only store hundreds of bits and have 5K-10K logic gates, but only 250-3K can be devoted to security tasks. Despite these restrictions, since the work of Sarma et. al [9] in 2002, most of the proposed solutions [1,2,15] are based on the use of hash functions. Although this apparently constitutes a good and secure solution, engineers face the non-trivial problem of implementing cryptographic hash functions with only between 250-3K gates. In most of the proposals, no explicit algorithms are suggested and finding one is not an easy issue since traditional hash functions (MD5, SHA-1, SHA-2) cannot be used [11]. In [16] we find a recent work on the implementation of a new hash function with a reduced number

of gates, but although this proposal seems to be light enough to fit in a low-cost RFID tag, the security of this hash scheme remains as an open question.

The remainder of the paper is organized as follows. In Sect. 2, we propose an Efficient Mutual-Authentication Protocol (*EMAP*) for low-cost RFID tags. A security evaluation and performance analysis of this new protocol is presented in Sect. 3. In Sect. 4, the proposed architecture for implementing our protocol is explained in detail. Finally, concluding remarks appear in Sect. 5.

2 Efficient-Lightweight Protocol

Like other authors, we think that the security of low-cost RFID tags can be improved with *minimalist cryptography* [5,12]. Following this direction, an extremely efficient lightweight mutual-authentication protocol, named EMAP, is proposed in this paper.

2.1 Suppositions of the Model

Our protocol is based on the use of pseudonyms, concretely on *index-pseudonyms* (*IDSs*). An *index-pseudonym* (96-bit length) is the index of a table (a row) where all the information about a tag is stored. Each tag has an associated key which is divided in four parts of 96 bits ($K = K1 \parallel K2 \parallel K3 \parallel K4$). As the *IDS* and the key (K) need to be updated, we need 480 bits of rewritable memory (EEPROM or FRAM) in total. A ROM memory to store the 96-bit static tag identification number (*ID*) is also required.

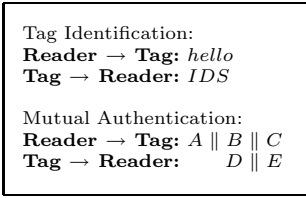
Costly operations such as random number generation will be done by readers. On the contrary, as tags are very limited devices that only have less than 1K logic gates for security functions, only simple operations are available: bitwise xor (\oplus), bitwise and (\wedge), and bitwise or (\vee). Multiplication have not been included because is a very costly operation [6].

Due to the fact that most low-cost tags are passive, the communication must be initiated by readers. We also suppose that both the backward and the forward channel can be listened by an attacker. Finally, we assume that the communication channel between the reader and the database is secure.

2.2 The Protocol

We can split our protocol proposal in four main stages: tag identification, mutual authentication, index-pseudonym updating, and key updating. In this section, we outline how the protocol works, while in the next one a security and performance analysis is presented.

Tag Identification. Before starting the protocol for mutual authentication, the reader should identify the tag. The reader will send a *hello* message to the tag, which answers by sending its current *index-pseudonym* (*IDS*). By means of this *IDS*, the reader will be able to access to the secret key of the tag ($K = K1 \parallel K2 \parallel K3 \parallel K4$), which is necessary to carry out the next authentication stage.



$$A = IDS_{tag(i)}^{(n)} \oplus K1_{tag(i)}^{(n)} \oplus n1 \tag{1}$$

$$B = (IDS_{tag(i)}^{(n)} \vee K2_{tag(i)}^{(n)}) \oplus n1 \tag{2}$$

$$C = IDS_{tag(i)}^{(n)} \oplus K3_{tag(i)}^{(n)} \oplus n2 \tag{3}$$

$$D = (IDS_{tag(i)}^{(n)} \wedge K4_{tag(i)}^{(n)}) \oplus n2 \tag{4}$$

$$E = (IDS_{tag(i)}^{(n)} \wedge n1 \vee n2) \oplus ID_{tag(i)} \bigoplus_{I=1}^4 KI_{tag(i)}^{(n)} \tag{5}$$

Fig. 1. EMAP Protocol

Mutual Authentication. Our protocol consists in the exchange of two messages between the reader and the tag. An execution of the protocol is shown in *Figure 1*. The $\bigoplus_{I=1}^N$ operation represents an N-elements addition with the bitwise xor operator ($K1 \oplus K2 \oplus \dots \oplus KN$).

- Reader Authentication: The reader will generate two random numbers $n1$ and $n2$. With $n1$ and the subkeys $K1$ and $K2$, the reader will generate the submessages A and B . With $n2$ and $K3$, it will generate the submessage C .
- Tag Authentication: With the submessages A and B , the tag will authenticate the reader and obtain $n1$. From the submessage C , the tag will obtain the random number $n2$, that will be used in the index-pseudonym and key updating. Once these verifications are performed, the tag will generate the answer message. This message will be composed of two parts D and E . The submessage D will allow to authenticate the tag and by means of E its static identifier will be transmitted in a secure form.

We have analyzed the statistical properties of these five submessages with three well-known suites of randomness tests, namely ENT [13], DIEHARD [7] and NIST [10]: we have generated a 300MB-file for every message. Due to extension restrictions the reports are not shown in the paper.¹ The results point to ensure submessages are not easily distinguishable from a random source, not even for the eavesdropper/cryptanalyst. As we can verify in *Equation 5*, submessage E uses more operations than the rest. We have put particular emphasis on the properties of submessage E due to the fact that in it the tag sends its more valuable information: the static identification number (ID).

Pseudonym Index Updating. Once the tag and the reader have mutually authenticated, each one has to update the index-pseudonym.

$$IDS_{tag(i)}^{(n+1)} = IDS_{tag(i)}^{(n)} \oplus n2 \oplus K1_{tag(i)}^{(n)} \tag{6}$$

The statistical properties of this sequence is good owing to the use of an xor with a random number ($n2$). In connection with the speed requirements, we have only used three basic operations (bitwise xor).

¹ The whole reports are available in <http://163.117.149.208/emap/>

Key Updating. The key updating will be carry out, as will the index-pseudonym updating, after the mutual authentication. As tags are very computationally constrained devices, this task should be made only by using efficient operations: bitwise xor (\oplus), bitwise and (\wedge), and bitwise or (\vee). These operations have already been implemented in the tag for the normal protocol running, so its use will not imply an increase in the gate counting. In order to improve the security of the key updating algorithm, a parity function will be used.² Nevertheless, the speed requirements of tags should be kept in mind; a tag must be able to answer 50 times/sec (see Sect. 4). These speed requirements put a limit on the number of operations that can be performed with each component of the key (KI). Taking all these considerations into account, the proposed equations for key updating are the following ones:

$$K1_{tag(i)}^{(n+1)} = K1_{tag(i)}^{(n)} \oplus n2 \oplus (ID_{tag(i)}(1 : 48) || F_p(K4_{tag(i)}^{(n)}) || F_p(K3_{tag(i)}^{(n)})) \quad (7)$$

$$K2_{tag(i)}^{(n+1)} = K2_{tag(i)}^{(n)} \oplus n2 \oplus (F_p(K1_{tag(i)}^{(n)}) || F_p(K4_{tag(i)}^{(n)}) || ID_{tag(i)}(49 : 96)) \quad (8)$$

$$K3_{tag(i)}^{(n+1)} = K3_{tag(i)}^{(n)} \oplus n1 \oplus (ID_{tag(i)}(1 : 48) || F_p(K4_{tag(i)}^{(n)}) || F_p(K2_{tag(i)}^{(n)})) \quad (9)$$

$$K4_{tag(i)}^{(n+1)} = K4_{tag(i)}^{(n)} \oplus n1 \oplus (F_p(K3_{tag(i)}^{(n)}) || F_p(K1_{tag(i)}^{(n)}) || ID_{tag(i)}(49 : 96)) \quad (10)$$

The statistical properties of these four sequences are good because of in each sequence there is an xor with a random number ($n1$ or $n2$). According to the speed requirements, for the worst case, which is obtained on the 8 bit architecture, a tag can authenticate 89 times per second, so we are able to successfully fulfill the speed requirements in all cases (see Sect. 4).

3 Evaluation

3.1 Security Analysis

Once we have presented the proposed mutual-authentication protocol, we will evaluate its security, studying the same properties that Yang analyzes in [15].

1. *User Data Confidentiality*

The tag ID must be kept secure to guarantee user privacy. The tag sends in the message E ($E = (IDS_{tag(i)}^{(n)} \wedge n1 \vee n2) \oplus ID_{tag(i)} \oplus_{I=1}^4 KI_{tag(i)}^{(n)}$) hiding the tag ID to a nearby eavesdropper equipped with an RFID reader.

2. *Tag Anonymity*

As the ID of the tag is static, we should send it, and all other interchanged messages in seemingly random wraps (i.e. to an eavesdropper, random numbers are sent). As we have seen, readers generate the message ($A||B||C$).

This message will serve to authenticate him, as well as to transmit in a

² Parity function ($F_p(X)$): The 96-bit number X is divided in twenty four 4-bit blocks. For each block we obtain a parity bit, getting 24 parity bits. See Sect. 4 for more details.

secure form the random numbers $n1$ and $n2$ to the tag. This two random numbers ($n1, n2$) will be used to hide the tag *ID* as well as to update the *index-pseudonym* and the associated key. By means of this mechanism we are able to make almost all the computational load to fall on the side of RFID readers, since one of our hypothesis is that very low-cost tags can not generate random numbers. Thus, tag anonymity is guaranteed and the location privacy of a tag owner is not compromised either.

There is one interesting scenario that we will explain with more detail in the following, as one could think that in this case, the tracking of a tag owner is possible. In this scenario, the attacker sends *hello* messages to the tag and receives as answer the *IDS* from it. Then, he stops the authentication step. A little time later he repeats the process, hoping that the *IDS* has not changed yet. We know that if the authentication process failed, the *IDS* can not be updated. The attacker can not generally track the owner tag because it is very probable that between two successive requests of the attacker, the tag is read by one or several legitimate readers, who will update the *IDS*. If an intruder wants to guarantee that the *IDS* has not changed, it needs to send more than 50 answers/sec in order to saturate the tag, so not allowing a legitimate reader to access it. In this case, this attack would be considered a DoS attack, which is an inherent problem in RFID technology as it happens in other technologies that use the radio channel. Unfortunately, for the moment, there is no known solution for it (instead of spread spectrum).

3. **Data Integrity**

A part of the memory of the tag is rewritable, so modifications are possible. In this part of the memory, the tag stores the *index-pseudonym* and the key associated with itself. If an attacker does succeed in modifying this part of the memory, then the reader would not recognize the tag and should implement the updating protocol of the database.

4. **Mutual Authentication**

We have designed the protocol with both reader-to-tag authentication (message $A \parallel B \parallel C$), and tag-to-reader authentication (message $D \parallel E$).

5. **Forward Security**

Forward security is the property that privacy of messages sent today will be valid tomorrow [8]. Since key updating is fulfilled after the mutual authentication, a future security compromise on an RFID tag will not reveal data previously transmitted.

6. **Man-in-the-middle Attack Prevention**

A man-in-the-middle attack is not possible because our proposal is based on a mutual authentication, in which two random numbers ($n1, n2$), refreshed with each iteration of the protocol, are used.

7. **Replay Attack Prevention**

An eavesdropper could store all the messages interchanged between the reader and the tag (different protocol runs). Then, he can try to impersonate a reader, re-sending the message ($A \parallel B \parallel C$) seen in any of the protocol runs. It seems that this could cause the losing of synchronization

Table 1. Comparison Between Protocols

Protocol	HLS [14]	EHLS [14]	HBVI [4]	MAP [15]	EMAP
User Data Confidentiality	×	△	△	○	○
Tag Anonymity	×	△	△	○	○
Data Integrity	△	△	○	○	△
Mutual Authentication	△	△	△	○	○
Forward Security	△	△	○	○	○
Man-in-the-middle Attack Prevention	△	△	×	○	○
Replay Attack Prevention	△	△	○	○	○
Forgery Resistance	×	×	×	○	○
Data Recovery	×	×	○	○	×

†† Notation: ○ Satisfied △ Partially satisfied × Not Satisfied

between the database and the tag, but this is not the case because after the mutual authentication, the *index-pseudonym* (*IDS*) and the key K ($K = K1 \parallel K2 \parallel K3 \parallel K4$) were updated.

8. **Forgery Resistance**

The information stored in the tag is sent operated (bitwise xor (\oplus), bitwise and (\wedge), and bitwise or (\vee)) with random numbers ($n1, n2$). Therefore the simple copy of information of the tag by eavesdropping is not possible.

9. **Data Recovery**

Intercepting or blocking of messages is a denial-of-service attack preventing tag identification. As we do not consider that these attacks can be a serious problem for very low-cost RFID tags, our protocol does not particularly focus on providing data recovery.

In those scenarios in which this problem is considered important, an extended version of the protocol is possible and quite straightforward. In this implementation each tag will have $l + 1$ database records, the first one associated with the actual *index-pseudonym* (n) and the others associated with the potential next *index-pseudonyms* ($n + 1, \dots, n + l$). Moreover, each tag will need k bits additionally of ROM memory to store the Associated Data Base Entry like in [4]. As before, the reader will use the *IDS* to access all the information associated with the tag. The reader will store a potential *IDS* each time the answer of the tag is blocked (uncertainty state). Once the tag and the reader have been authenticated mutually, the potential *IDS* could be deleted (synchronized state). The storage of the potential *IDS* will allow to easily recover from the lose or interception of messages.

Table 1 shows a comparison of the security requirements made by Yang [15], as met by different proposals in the literature. We have added our proposal (EMAP) in the last column.

3.2 Performance Analysis

Before evaluate the security of the protocol a performance analysis will be presented (see Table 2), considering the following overheads (computation, storage, and communication) as in Yang [15].

Table 2. Computational Loads and Required Memory

Protocol	Entity	HLS [14]	EHLS [14]	HBVI [4]	MAP [15]	EMAP
No. of Hash Operation	T	1	2	3	2	\neg
No. of Keyed Hash Operation	B	\neg	Nt	3	2Nt	\neg
No. of RGN Operation	R	\neg	\neg	\neg	1	\neg
	B	\neg	\neg	\neg	1	\neg
	T	\neg	1	\neg	\neg	\neg
	R	\neg	\neg	\neg	1	\neg
	B	\neg	\neg	1	\neg	\neg
No. of Basic Operation ^{1,2}	T	\neg	\neg	\neg	4	22
	$R+B$	\neg	\neg	\neg	$2(Nt+1)$	25
No. of Encryption	B	\neg	\neg	\neg	1	\neg
No. of Decryption	R	\neg	\neg	\neg	1	\neg
Number of Authentication Steps		6	5	5	5	4
Required	T	$1\frac{1}{2}L$	$1L$	$3L$	$2\frac{1}{2}L4$	$6L$
Memory Size	$R+B$	$2\frac{1}{2}L$	$1\frac{1}{2}L$	$9L$	$9\frac{1}{2}L$	$6L$

†† Notation: \neg : Not require Nt: Number of Tags L : Size of Required Memory
¹Basic Operations: Bitwise xor (\oplus), Bitwise and (\wedge), and Bitwise or (\vee)
²Parity function has been included as a basic operation

1. **Computation Overhead**

Low-cost RFID tags are very limited devices, with only a small amounts of memory, and very constrained computationally (<1K logic gates to security-related tasks). Additionally, one of the main drawbacks that hash-based solutions have is that the load on the server side (R+B) is proportional to the number of tags, as it happens in Yang’s solution [15]. Our proposal (EMAP) have completely solved this problem by using an *index-pseudonym*.

2. **Storage Overhead**

As Yang does, we assume that all components are L -bits sized, that the RNG and the hash function are $h, h_k : \{0, 1\}^* \rightarrow \{0, 1\}^{\frac{1}{2}L}$ and $r \in_U \{0, 1\}^L$. As we see in Sect. 2.1 each tag has to store an L -bit *index-pseudonym* (IDS) and an associate key (K) of four L -bit components. Moreover, the tag has to store an unique L -bit identification number (ID). The reader has to store the same information, so it requires a memory of $6L$ bits.

3. **Communication Overhead**

As we can see in *Table 2*, according the number of interchanged messages to accomplish mutual authentication tag-reader, our protocol is the most efficient. As low cost tags are passive and that the communication can only be initiated by a reader, four rounds may be considered as a reasonable number of rounds for mutual authentication in RFID environments.

4 **Implementation**

In this section, we will explain in detail the proposed architecture for implementing our protocol. The proposed architecture is independent of the word length used. We have analyzed the features of four different word length ($m = 8, 24, 48, 96$ bits). In *Figure 2* we can see a scheme of the proposed architecture. On the left of the figure we have the memory, which is filled with the *index-pseudonym*

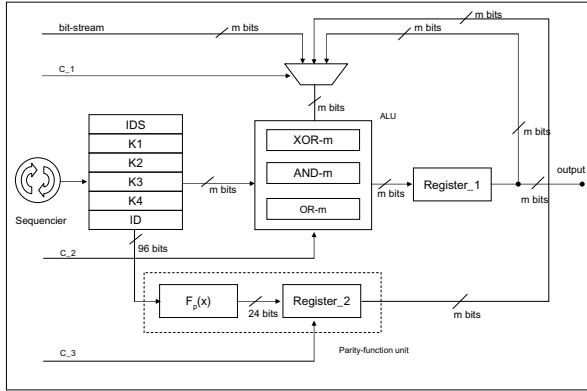


Fig. 2. Logic Scheme

(IDS), the key K ($K1 \parallel K2 \parallel K3 \parallel K4$), and the static identification number (ID). The access to the memory is controlled by a sequencer. Due to the fact that messages are build up of three or more components, we will need a m -bit register to store intermediate results. In the middle of the figure we have the Arithmetic Logic Unit (ALU). This unit will make the following m -bit operations: bitwise xor (\oplus), bitwise and (\wedge), and bitwise or (\vee). The ALU has two inputs, one of these values stored in the memory and another which is selected (c_1) between one of these three values: the bitstream, the value stored in the register_1 and the the result of the parity-function unit. The control signal c_2 will select the operation that will be used in the ALU. At the bottom of the figure we can see the parity-function unit. This unit will be used each time the key is updated, in particular twice with each part of K ($K1 \parallel K2 \parallel K3 \parallel K4$). In order to carry out the temporal requirements we have decided to implement this function in just one block. This function has an input length of 96 bits and a 24-bits output. The input is divided in blocks of 4 bits, which are processed to obtain an output bit. For example, for the first four bits, x_1 is xored with x_2 , x_3 is xored with x_4 , and finally the corresponding outputs are xored ($(x_1 \oplus x_2) \oplus (x_3 \oplus x_4)$). So for 96 bits of input, we need 72 logical gates (24×3) for implementing the parity function. The output of this function is stored in the register_2 of dimension m ($m = 24, 48, 96$ bits). The control signal c_3 will select when a 24-bit shift has to be done in the register.

It is a common assumption that a maximum of 50 tags can be authenticated per second. As in [3], due to the low-power restrictions of RFID tags, the clock frequency must be set to 100 KHz. So, a tag may use up to 2000 clock cycles to answer a reader. In the worse case of our protocol ($m = 8$ bits), we need 1120 clock cycles for running the protocol (mutual authentication, pseudonym updating, and key updating). So, if we consider that the clock frequency is set to 100KHz [3], this means that the tag answers in 11.2 milliseconds. A tag can authenticate 89 times per second, so the temporary requirements are fulfilled in all the cases.

Another important aspect to study is the number of logical gates necessary for implementing the proposed protocol. The functions bitwise xor (\oplus), bitwise and (\wedge), and bitwise or (\vee) will be implemented with the same number of logic gates like the word length (m). As seen above, 72 logical gates will be needed for implementing the parity function. Additionally, an extra 30% of logic gates are added up for control functions. In the worst case ($m = 96$ bits) the protocol only needs around 500 gates. Moreover, although we have not implemented the circuit physically, due to the known fact that power consumption and circuit area are proportional to the number of logical gates, it seems that our implementation will be suitable even for very low-cost RFID tags.

Table 3. Features

Word's length		8-bits	24-bits	48-bits	96-bits
Number of. Gates	ALU	24	72	144	288
	Parity Function	72	72	72	72
	Control	29	43	65	108
	Total	125	187	281	468
Number of Clock Cycles		1120	416	240	152
Answer/sec		89	240	417	658

5 Conclusions

RFIDs tags are devices limited to hundreds of bits of store, and with roughly 250-3K gates devoted to security-related tasks. Cryptographic primitives such as PRNGs, block ciphers, and hash functions lie well beyond the computational reach of very low cost RFID tags, but until now, most of the security solutions for RFID are based on those. A new approach must be taken to tackle the problem, at least for low-cost RFID tags. For this reason, we propose an extremely efficient lightweight mutual-authentication protocol (*EMAP*) that could be implemented in low-cost tags (<1K logic gates). In order to be able to use our proposal, tags should be fitted with a small portion of rewritable memory (EEPROM or FRAM) and another read-only memory (ROM). The assumption of having access to rewritable memory is also made in all the existing solutions based on hash functions.

In spite of being very limited in resources, the main security aspects of RFID systems (privacy, tracking) have been consider in this article and solved efficiently (less than 500 gates are needed even in the worst implementation, in our case $m = 96$ bits). As shown in *Table 2*, our protocol displays superior benefits to many of the solutions based on hash functions. So, not only we have been able to avoid the privacy and tracking problems, but also many other attacks such as the man-in-the-middle attack, replay attack, etc.

Finally, another paramount characteristic of our scheme is its efficiency: tag identification by a valid reader do not require exhaustive search in the back-end database. Furthermore, only two messages need to be exchanged in the identification stage and another two in the mutual authentication stage.

References

1. E.Y. Choi, S.M. Lee, and D.H. Lee. Efficient RFID authentication protocol for ubiquitous computing environment. In *Proc. of SECUBIQ'05*, 2005.
2. T. Dimitriou. A lightweight RFID protocol to protect against traceability and cloning attacks. In *Proc. of SECURECOMM'05*, 2005.
3. M. Feldhofer, S. Dominikus, and J. Wolkerstorfer. Strong authentication for RFID systems using the AES algorithm. In *Proc. of CHES'04*, volume 3156 of *LNCS*, pages 357–370, 2004.
4. D. Henrici and P. Müller. Hash-based enhancement of location privacy for radio-frequency identification devices using varying identifiers. In *Proc. of PERSEC'04*, pages 149–153. IEEE Computer Society, 2004.
5. A. Juels. Minimalist cryptography for low-cost RFID tags. In *Proc. of SCN'04*, volume 3352 of *LNCS*, pages 149–164. Springer-Verlag, 2004.
6. T. Lohmann, M. Schneider, and C. Ruland. Analysis of power constraints for cryptographic algorithms in mid-cost RFID tags. In *Proc. of CARDIS'06*, volume 3928 of *LNCS*, pages 278–288, 2006.
7. G. Marsaglia and W.W. Tsang. Some difficult-to-pass tests of randomness. *Journal of Statistical Software*, Volume 7, Issue 3:37–51, 2002.
8. M. Ohkubo, K. Suzuki, and S. Kinoshita. Cryptographic approach to “privacy-friendly” tags. In *RFID Privacy Workshop*, 2003.
9. S.E. Sarma, S.A. Weis, and D.W. Engels. RFID Systems and Security and Privacy Implications. In *Proc. of CHES'02*, volume 2523, pages 454–470. LNCS, 2002.
10. C. Suresh, Charanjit J., J.R. Rao, and P. Rohatgi. A cautionary note regarding evaluation of AES candidates on smart-cards. In *Second Advanced Encryption Standard (AES) Candidate Conference*. <http://csrc.nist.gov/encryption/aes/round1/conf2/aes2conf.htm>, 1999.
11. Datasheet Helion Technology. MD5, SHA-1, SHA-256 hash core for Asic. <http://www.heliontech.com>, 2005.
12. I. Vajda and L. Buttyán. Lightweight authentication protocols for low-cost RFID tags. In *Proc. of UBICOMP'03*, 2003.
13. J. Walker. ENT Randomness Test. <http://www.fourmilab.ch/random/>, 1998.
14. S.A. Weis, S.E. Sarma, R.L. Rivest, and D.W. Engels. Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems. In *Security in Pervasive Comp.*, volume 2802 of *LNCS*, pages 201–212, 2004.
15. J. Yang, J. Park, H. Lee, K. Ren, and K. Kim. Mutual authentication protocol for low-cost RFID. Ecrypt Workshop on RFID and Lightweight Crypto, 2005.
16. K. Yksel, J.P. Kaps, and B. Sunar. Universal hash functions for emerging ultra-low-power networks. In *Proc. of CNDS'04*, 2004.

Secure EPCglobal Class-1 Gen-2 RFID System Against Security and Privacy Problems*

Kyoung Hyun Kim¹, Eun Young Choi², Su Mi Lee³, and Dong Hoon Lee⁴

Center for Information Security Technologies(CIST),
Korea University, 1, 5-Ka, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
visionkkh@korea.ac.kr¹, bluecey@cist.korea.ac.kr², smlee@cist.korea.ac.kr³,
donghlee@korea.ac.kr⁴

Abstract. Radio Frequency Identification (RFID) system is an important technology in ubiquitous computing environment. RFID system should be compatible with most RFID system applications to support the ubiquitous computing environment. Recently, researchers had studied about RFID standardization. After all, EPCglobal Class-1 Gen-2 (C1G2) RFID is selected as an international standard of RFID systems. Unfortunately, it has fatal security problems to be vulnerable to information leakage and traceability since a tag of EPCglobal C1G2 emits its fixed ID(EPC) without hiding or modifying. A main goal of our work is to propose the secure protocol well suitable for EPCglobal C1G2. First of all, our protocol exactly follows RFID standard with only current capabilities of a tag approved in the standard, assuring that our protocol is secure against impersonation, information leakage, and traceability etc.

1 Introduction

Radio Frequency Identification (RFID) is an automatic identification system using a microchip that has capability of transmitting a unique serial number or other additional data through RF(radio frequency) signals. The advantage of RFID system is to provide good properties; many-to-many communication, wireless data transmission, and self-computing. These benefits enable more wider range of application fields such as manufacturing, supply chain management, and inventory control, etc. For the reason, RFID technology is expected to be massively deployed in the near future.

To popularize RFID system, several organizations including EPCglobal and ISO had been working on standardization. In particular, the EPCglobal Class-1 Gen-2 (C1G2) RFID was adopted as an international standard by ISO/IEC. However, the EPCglobal C1G2 specification has fatal security problems. Generally, RF signals make RFID system vulnerable to various attacks such as eavesdropping, traffic analysis, spoofing and denial of service. These attacks may disclose sensitive information of tags and hence infringe on a user's privacy. Another type of privacy violation is traceability which establishes a relation between a user and a tag. If a link can be established between a user and the

* This research was supported by the Seoul R&BD Program(10665), Korea.

tag he/she holds, the tracing of the tag makes the tracing of the user possible [4]. Nevertheless the basic security concerns have been already considered in the related literatures, they are not considered in EPCglobal C1G2 specification. Namely, a tag emits EPC without hiding or modifying and anybody can be readable it. An adversary can easily obtain EPC by eavesdropping on communication between a reader and a tag. This can cause user's privacy infringement, because anybody can access a service which provides information of the object corresponding EPC [14]. Furthermore, the adversary with eavesdropping messages can clearly trace a user's movement, since EPC is a unchanging value. These problems in the RFID standard will make a limit of standard's public usage because security and privacy issue is very sensitive for people.

For providing security in the standard, researchers should try to modify steps of the standard or add new components to the standard. The additional works have to pay a lot of expense and may be a bar to popularize RFID system through the standard. Therefore, it is important work to provide various security considered in general RFID system without modifying of the standard. We propose the secure protocol against security and privacy problems which maintains each step and components of the standard.

1.1 Related Works and Contributions

Related works. Researchers have recognized the security problems of tags [2]. We describe related studies below. The simplest physical approach is to *Kill* tags [7]. *Kill* technique is to restrict the use of a tag by removing its ID. However, this method is not a useful solution. A tag will be used as active state in numerous applications. For example, in animal tracking system, the tag should not be killed, because the tags must always be in working state.

Another approach is using hash function [3,7,9]. This approach can prevent an exposure of tag's ID using one-wayness property of a hash function. However, implementing a standard cryptographic hash function, such as MD5 or SHA-1, in a tag is beyond the capabilities of today's tags. So, current EPCglobal C1G2 specification does not make use of hash function.

Recently, Juels *et al.* considered operation complexity of tag [8]. The protocol does not require for a tag to perform any cryptographic operations except XOR, simple bit-operation. It may be possible to implement the protocol in current low-cost RFID systems. Unfortunately, it did not consider eavesdropping and privacy issues and is not secure against privacy and information leakage.

Thereafter, Duc *et al.* proposed a scheme for EPCglobal C1G2 [11]. They reported the security problems of EPCglobal C1G2, and suggested the solution about the problems. However, we had to modify composition of the standard for applying the scheme in the standard, since the authors of the paper [11] do not consider working steps of EPCglobal C1G2. In addition, the scheme can not re-synchronize automatically when synchronization is broken. The works require additional costs, comparing with schemes which does not need re-synchronization.

Contribution. In this paper, we analyze working steps and security threats of EPCglobal C1G2 in detail. This will help several security researches of

EPCglobal C1G2. We propose a secure protocol which is suitable for EPCglobal C1G2. The main contribution of our paper is that our protocol can apply to EPCglobal C1G2 without any modifying steps or components of the standard and is secure against security problems. In addition, our protocol does not need synchronization. By the advantages of our paper, the standard reduce entire costs adding security as well as is spread out wide and fast in our life.

Organization of the paper. This paper is organized as follows: In Section 2, we describe RFID systems and analyze security problems of RFID system. In Section 3, We analyze EPCglobal C1G2 and threats of EPCglobal C1G2. Then We describe our proposed protocol in Section 4. We analyze our protocol in security in Section 5. Finally, we conclude in Section 6.

2 RFID System

2.1 Components

In general, RFID system consists of a tag, a reader, and a back-end server.

- **Tag:** A tag is a small and low-priced chip which is adhered on objects. It consists of only a microchip with limited functionality and data storage, and an antenna to wireless communication with reading devices. RFID tags can be classified into two types, active or passive depending on powering technique. While an active tag can generate power by itself, a passive tag is not able to supply a power by itself. Therefore the passive tag obtains power from the reading devices when it is within range of some reading devices.
- **Reader:** A reader can read and re-write the data in a tag. A reader queries a tag to obtain the tag's contents though RF interface. After the reader queries to a tag and receives some information from the tag, the reader forwards the information to a back-end server.
- **Back-end server:** A back-end server is a device that manages and stores various information such as EPC for each tag. It can also determine a tag's identity from the information of a tag sent by an authenticated reader.

2.2 Threats of RFID System

In RFID systems, because of wireless communication between a reader and a tag, an adversary can monitor all messages transmitted between a reader and a tag. The adversary can also infringe upon a user's privacy using various methods. Therefore, RFID systems must be designed to be secure against attacks such as eavesdropping and impersonation.

- **Eavesdropping:** A passive adversary can eavesdrop on messages between a reader and a tag. By eavesdropping, she may obtain a user's secret information. Therefore, RFID systems should be considered that she cannot get any secret information from the eavesdropped messages.
- **Impersonation:** An active adversary can query to a tag and a reader in RFID systems. By this property, she can impersonate the target tag or the legitimate reader. When a target tag communicates with a legitimate

reader, an adversary can collect the messages sent to the reader from the tag. With the message, the adversary makes a clone tag in which information of a target tag is stored. When the legitimate reader sends a query, the clone tag can reply the message in response, using the information of a target tag. Then the legitimate reader may consider the clone tag as a legitimate one.

- **Information Leakage:** If RFID systems are used widely, users will have various tagged objects. Some of objects such as expensive products and medicine provide quite personal and sensitive information that the user does not want anyone to know. If RFID systems are not designed to protect information of tag, user's information can be leaked without acknowledgment of the user.
- **Traceability:** When a user has special tagged objects, an adversary can trace user's movement using messages transmitted by the tags. In the concrete, when a target tag transmits a response to a nearby reader, an adversary can record the transmitted message and can establish a link between the response and the target tag. Once a link established, the adversary is able to know the user's location history.

3 EPCglobal Class-1 Gen-2 RFID System

In this section, we explain EPCglobal Class-1 Gen-2 RFID system[13], and analyze security problems of it.

3.1 EPCglobal Class-1 Gen-2 RFID

EPCglobal C1G2 was adopted as 18000-6 international Standard by ISO/IEC. As result, RFID system will be able to be recognized without confusion. EPCglobal C1G2 tag has properties as follows [11,13]:

1. Tag is passive.
2. Tag uses UHF band (800-960 MHz) and communication range is 2-10m.
3. Tag has on-chip Pseudo-Random Number Generator (PRNG) and Cyclic Redundancy Code (CRC).
4. Tag has two 32-bit PIN for kill command and access command. The kill PIN is used to kill the tag. The access PIN is used to write into the tag or to read something in password fields.

EPCglobal C1G2 operates as shown in Fig.1 and a detailed description of each step is as follows:

- (1) A reader sends a request message to a tag.
- (2) Each tag which is received the request generates a 16-bit random value (RN16) using Pseudo-Random Number Generator. After that, each tag inputs the random value(RN16) into a slot counter, and starts a slot counter. A slot counter decreases the random value as a regular interval. When a slot counter becomes zero, the tag sends the random value(RN16) to the reader.
- (3) As response, the reader sends ACK which has the random value(RN16) in reserved field.

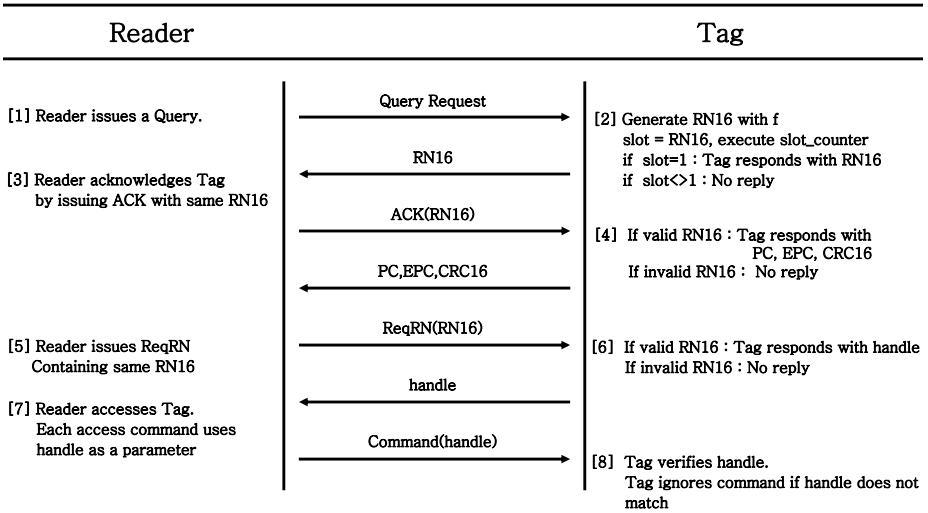


Fig. 1. Process of EPCglobal Class-1 Gen-2 RFID

- (4) The tag which is received ACK compares a random value in ACK with RN16 of the tag. If the values are the same, the tag sends PC(Protocol-Control), EPC(Electronic Product Code), and CRC to the reader. The PC bits contain Physical-layer information.

In addition, the reader can make use of memory-writing command or kill command, as follows after above steps:

- (5) The reader sends ReqRN which contains RN16 to the tag.
- (6) The tag which receives ReqRN compares a random value in ReqRN with RN16 of the tag. If the two values are the same, the tag passes handle to the reader.
- (7) If the reader gets handle of the tag, it can construct memory-writing or kill command. The reader sends PIN for command of access or kill.
- (8) The tag verifies PIN which is received from the reader. If the PIN is right, the tag carries out command. When the reader sends PIN, PIN is XORed with RN16 as RN16 is repeated twice.

In the EPCglobal C1G2 specification, a RFID tag is capable of generating 16-bit pseudo-random [13]. The pseudo-random number is not used for security but used for making a new session between a reader and a tag. A reader signal can wake up several tags. If many tags act upon reader signal simultaneously, then they may have a collision. By the reason, it needs the system of collision avoidance. The pseudo-random number is used for distinguishing a tag from several tags and preventing the collision.

3.2 Threats of EPCglobal Class-1 Gen-2 RFID System

According to EPCglobal C1G2 specification, whenever a tag authenticates a reader, the tag emits EPC as plain text. With only eavesdropping on communication

between a reader and a tag, an adversary can obtain EPC of the tag. It can bring about security problems as follows:

- **Impersonation:** An active adversary easily impersonates legitimate tags. The adversary stores EPC of a legitimate tag by eavesdropping. After a legitimate reader sends a request message, the adversary emits a stored value (legitimate EPC). Then the reader may consider the value as legitimate.
- **Information Leakage:** If an adversary can know EPC of user's tag, the adversary will become aware of some information of user's object, because EPC discovery service [14] is publicized. Therefore it can cause privacy infringement of the user.
- **Traceability:** Because tags emit fixed EPC, an adversary with EPC in user's tag can trace the user's movement in EPCglobal C1G2. When tags transmits EPC to a reader, the adversary records EPCs by eavesdropping. Then the adversary can establish a link between the user's tag and its EPC. Once a link is established, the adversary can know user's movement history.

Another threat is that PIN can be disclose. When PIN is sent, PIN is XORed with random value (RN16). But RN16 is sent as plain text when session starts. Just by eavesdropping the 16-bit pseudo-random number and the XORed PIN, the PIN can be recovered by an adversary. If PIN is exposed, the adversary can write and delete the memory, and kill a tag.

4 Our Protocol Suitable for EPCglobal Class-1 Gen-2 RFID System

In this section, we propose a privacy protection protocol suitable for EPCglobal C1G2. First of all, we define notations. After that, we describe our protocol.

4.1 Our Protocol

The following notations are used for the computational operations to simplify the description.

Notation	Description
RT32	32-bit random number generated by a tag
RR32	32-bit random number generated by a reader
PIN1, PIN2	Two EPCglobal C1G2 PINs(access, kill)
EPC	Electronic Product Code
f	32-bit pseudo-random number generator
n	The number of tags in the system
\parallel	Concatenation of two inputs
\oplus	Exclusive-or of two inputs

We assume that the channel between a reader and a back-end server is secure. The channel between a tag and a reader is insecure because of wireless.

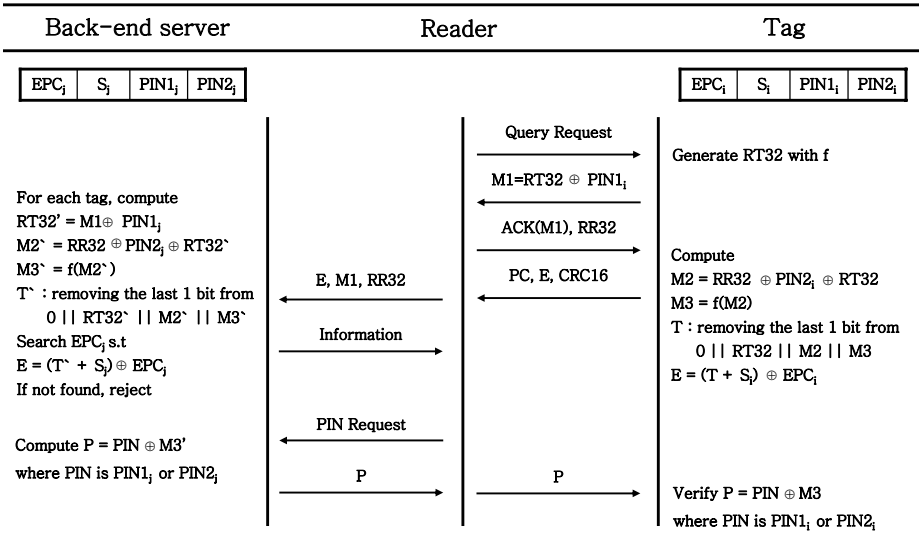


Fig. 2. Process of our proposed protocol

Our protocol exactly follows EPCglobal C1G2 steps with 96-bit EPC. The 16-bit pseudo-random number of EPCglobal C1G2 is not suitable for security. It is very small to use for security. Then we suggest that 32-bit pseudo-random number generator should be supported, in order to fulfill security and to take full advantage of 32-bit PIN. In this paper, we assume that both a reader and a tag use a 32-bit pseudo-random number generator in EPCglobal C1G2.

Each tag \mathcal{T}_i where $i \in \{1, \dots, n\}$, has EPC_i , $PIN1_i$, $PIN2_i$, and 96-bit secret value S_i of which the first bit is 0. Back-end server stores EPC_j , $PIN1_j$, $PIN2_j$, and S_j where $j \in \{1, \dots, n\}$, for each tag. Fig.2 shows the process of the proposed protocol, and the following gives a detailed description of each step:

- (1) A reader sends Query Request to a tag \mathcal{T}_i .
- (2) The tag \mathcal{T}_i generates 32-bit random RT32 from f . And it computes $M1 = RT32 \oplus PIN1_i$, sends M1 to the reader.
- (3) The reader generates a new 32-bit random RR32. The reader sends ACK(M1) and RR32.
- (4) The tag computes as follows:
 - $M2 = RR32 \oplus PIN2_i \oplus RT32$, $M3 = f(M2)$
 - T is a bit string of which the last 1-bit is removed from $0 \parallel RT32 \parallel M2 \parallel M3$. T is 96-bit and the first bit of T is 0.
 - $E = (T + S_i) \oplus EPC_i$, where T and S are 96-bit of which the first bit is 0. We must consider the first bit(carry) of $T+S_i$. Then we set 0 in the first bit of T and S_i .

The tag sends PC, E, CRC to the reader.

- (5) The reader sends E, M1, RR32 to a back-end server.

- (6) The back-end server searches EPC for each j where $j \in 1, \dots, n$ as follows:
- For each $\langle \text{EPC}_j, S_j, \text{PIN1}_j, \text{PIN2}_j \rangle$, compute $\text{RT32}' = \text{M1} \oplus \text{PIN1}_j$, $\text{M2}' = \text{RR32} \oplus \text{PIN2}_j \oplus \text{RT32}'$ and $\text{M3}' = f(\text{M2}')$.
 - T' is a bit string of which the last 1-bit is removed from $0 \parallel \text{RT32}' \parallel \text{M2}' \parallel \text{M3}'$
 - Search EPC such that $\text{E} = (\text{T}' + S_j) \oplus \text{EPC}_j$

When the reader wants to command writing, delete, and kill, PIN must be used. For protecting PIN, we accomplish as follows:

- (7) The reader requests PIN to the back-end server.
- (8) The back-end server computes $\text{P} = \text{M3} \oplus \text{PIN}$, where PIN is PIN1_j or PIN2_j , and sends P to the reader
- (9) The reader sends P to the tag for a command.
- (10) The tag verifies $\text{P} \oplus \text{M3} = \text{PIN1}_i$ or PIN2_i , if it is true, the tag executes the command.

5 Analysis

In this section, we give a security analysis of our proposed protocol.

5.1 Security Analysis

We analyze the security of our protocol under section 2.2. as follows:

- **Eavesdropping and Information leakage:** A reader which is not authenticated by a back-end server does not obtain a tag's EPC although the reader sends requests to the tag. By eavesdropping, an adversary can obtain $\langle \text{E}, \text{M1}, \text{RR32} \rangle$. In order to recover EPC from E, the adversary must know $\text{T} + S_i$ of each session. S_i is a secret value for each tag \mathcal{T} and T is made by random values. Therefore, the adversary does not obtain information about $\text{T} + S_i$. The adversary must guess 96-bits of $\text{T} + S_i$. The probability is $\frac{1}{2^{96}}$. In spite of the fact that the first bit of T and S_i is 0, the probability that the first bit (carry) of $\text{T} + S_i$ is 1 equals 1/2. Therefore, the probability that the adversary obtain EPC by eavesdropping is negligible. Because EPC is hidden every session, the adversary does not obtain information of a user.

In format of EPC, the front bit string of EPC is fixed and public. And for each tag, we can be aware of fixed part of EPC. For example, in 96-bit EPC, the first eight bits of EPC are version and length of serial number. An adversary can also know the front bits of EPC. Then the adversary can obtain the front bits of $\text{T} + S_i$, too. Even if we assume that the adversary knows all bits of EPC, for getting the security value S_i , the adversary must guess T. The T is made of random value RT32 and RR32. Because the two values are passed by PIN1 and PIN2, the adversary does not acquire the information about RT32 and RR32. Then, the probability guessing T is $\frac{1}{2^{64}}$. Therefore, although the front bit string of EPC is fixed and public, S_i is secure.

- Impersonation and Reply attack: An active adversary may be able to reply for cheating the legitimate reader, using data which legitimate tags emit. In proposed protocol, the values emitted by a tag are changed every session since random values are used by the reader and the tag. Because the probability that the legitimate reader and tag use same random values over twice is very small, it is hard that the adversary succeeds the reply attack.
- Traceability: An adversary can trace location of tags, if response of tag is same or similar pattern for each session. In proposed protocol, the adversary can obtain $\langle E, M1, RR32 \rangle$ by eavesdropping. $M1$ is changed for each session, and can be considered as a random value. $RR32$ becomes a different value for each session, if the legitimate reader. E is also changed every session. Although a malicious reader sends requests to a tag by using same $RR32$, it is difficult that an adversary makes an accurate estimate of E .

6 Conclusion

In spite of EPCglobal C1G2 system has been international standard, it is analyzed that it is insecure. In order to protect user's privacy in EPCglobal C1G2 system, we propose a RFID protocol which is secure against impersonation, information leakage, and traceability. We consider working steps and components of EPCglobal C1G2 system. Then our protocol does not alter composition of EPCglobal C1G2. Our protocol uses only approved abilities of tags in EPCglobal C1G2 specification. For that reason, our protocol is suitable for the application using EPCglobal C1G2. Namely, our protocol is a good alternative mechanism which can insert security in the specification without a lot of modification.

References

1. A. Juels, R. L. Rivest and M. Szudlo. *The Blocker Tag: Selective Blocking of RFID tags for Consumer Privacy*. In the 8th ACM Conference on Computer and Communications Security, pp. 103-111, ACM Press, 2003.
2. S. E. Sarma, S. A. Weis and D. W. Engels. *Radio-frequency identification systems*. CHES'02, vol.2523 LNCS, pp.454-469, Springer-Verlag, 2002.
3. M. Ohkubo, K. Suxuki and S. Kinoshita. *Efficient Hash-Chain Based RFID Privacy Protection Scheme*. Ubcomp2004 workshop.
4. G. Avoine and P. OecGJJS04hslin. *A scalable and Provably Secure Hash-Based RFID Protocol*. In IEEE PerSec 2005, Kauai Island, Hawaii, March, 2005.
5. L. Su Mi, H. Young Ju, L. Dong Hoon and L. Jong In. *Efficient Authentication for Low-Cost RFID systems*. ICCSA05, vol. 3480 LNCS, pp.619-629, Springer-Verlag, 2005.
6. C. Eun Young, L. Su-Mi, L. Dong Hoon *Efficient RFID Authentication Protocol for Ubiquitous Computing Environment*. EUC Workshops, 2005.
7. S. A. Weis, S. E. Sarma, S. A. Weis and D. W. Engels. *Security and privacy Aspects of Low-Cost Radio Frequency Identification Systems*. First International Conference on Security in Pervasive Computing, 2003. <http://theory.lcs.mit.edu/sweis/spc-rfid.pdf>

8. Ari Juels, *Strengthening EPC Tag against Cloning*. To Appear in the Proceedings of WiSe'05.
9. D. Henrici and P. Muller. *Hash-based Enhancement of Location Privacy for Radio-Frequency Identification Devices using Varying Identifiers*. PerSec'04 at IEEE Per-Com. 2004
10. Miyako Ohkubo, Koutarou Suzuki, and Shingo Kinoshita, *Efficient Hash-Chain Based RFID Privacy Protection Scheme*. In the Proceedings of International Conference on Ubiquitous Computing, Workshop Privacy, 2004.
11. Dang Nguyen Duc, Jaemin Park, Hyunrok Lee, and Kwangjo Kim, *Enhancing Security of EPCglobal Gen-2 RFID Tag against Traceability and Cloning* The Symposium on Cryptography and Information Security, 2006.
12. EPCglobal Inc., <http://www.epcglobalinc.org/>.
13. EPCglobal Inc., *Class 1 Generation 2 UHF RFID protocol for communication at 860Mhz-960Mhz version 1.0.9*
14. EPCglobal Inc., *EPCglobal Object Name Service (ONS) 1.0*

A Case Against Currently Used Hash Functions in RFID Protocols*

Martin Feldhofer and Christian Rechberger

Graz University of Technology
Institute for Applied Information Processing and Communications
Inffeldgasse 16a, A-8010 Graz, Austria
{Martin.Feldhofer, Christian.Rechberger}@iaik.tugraz.at

Abstract. Designers of RFID security protocols can choose between a wide variety of cryptographic algorithms. However, when implementing these algorithms on RFID tags fierce constraints have to be considered. Looking at the common assumption in the literature that hash functions are implementable in a manner suitable for RFID tags and thus heavily used by RFID security protocol designers we claim the following. Current standards and state-of-the-art low-power implementation techniques favor the use of block ciphers like the Advanced Encryption Standard (AES) instead of hash functions from the SHA family as building blocks for RFID security protocols. In turn, we present a low-power architecture for the widely recommended hash function SHA-256 which is the basis for the smallest and most energy-efficient ASIC implementation published so far. To back up our claim we compare the achieved results with the smallest available AES implementation. The AES module requires only a third of the chip area and half of the mean power. Our conclusions are even stronger since we can show that smaller hash functions like SHA-1, MD5 and MD4 are also less suitable for RFID tags than the AES. Our analysis of the reasons of this result gives some input for future hash function designs.

1 Introduction

In the last few years, many research activities were conducted in the area of RFID security. Various attacks on the used algorithms, protocols and implementations showed that the protection of RFID systems requires more attention. Many security protocols were proposed to protect the violation of privacy and the authenticity of goods. Most of them use symmetric cryptography because of the fierce constraints for RFID tag implementations. From the implementation point of view, the difficulties in RFID tag design are the very tight requirement for the production of RFID tags. In addition to low die-size requirements the power consumption of the RFID tag is of utmost importance.

In the HF frequency range at 13.56 MHz the maximum mean current consumption without reducing the operation range of the tags is 15 μ A. Due to the limited available chip area, the limited power consumption and the limited time, an algorithm is allowed

* This work origins from the Austrian Government funded project *SNAP* established under the embedded system program FIT-IT.

to execute, the selection of appropriate security algorithms and protocols is very important. Unfortunately, the use of public key cryptography is out of range with today's semiconductor process technologies. The required computational power cannot be included on RFID tags in terms of speed and power consumption. Therefore, primitives from symmetric cryptography are heavily used by protocol designers in this area.

Hash functions are conceptually simpler than block ciphers, since they do not need a key. In the RFID security community, it is commonly assumed that hash functions are therefore also the better choice from the implementation point of view. As a consequence, most of the proposed protocols for protecting RFID tags base on hash function implementations [1,3,7,9,13,16]. Only in the work of Feldhofer *et al.* [5] the use of block ciphers is discussed in more detail.

Our Contribution. In this paper we give conclusive evidence that it is better to use a standardized block cipher like AES [10] on RFID tags than using one of the standardized hash functions from the SHA family [11].

On the basis of a survey on existing RFID security protocols we see that either symmetric encryption primitives or hash functions are always the underlying building blocks if authentication is needed. In Section 2 we derive the requirements for the basic building blocks. In Section 3 we present the smallest and most power-efficient SHA-256 ASIC implementation published so far. In Section 4 we map these results to other MD4 family hash functions like SHA-1, MD5 and MD4 and we compare the achieved results of hash implementations with the low-power AES implementation of Feldhofer *et al.* [6] which requires less chip area and has lower power consumption figures. Based on that we derive two simple criteria that firstly explain the reason between the gap we observe, and secondly give hints for the design of new hash functions which are more suitable for RFID tags.

2 Cryptographic Primitives in RFID Security Protocols

The cryptographic literature offers a variety of cryptographic primitives which can be used as basic building blocks in the design of security protocols. Depending on assumptions about available keying material, difficulty and trust on underlying problems, or implementation constraints, a multitude of options is available. In the context of RFID systems, the choice is somewhat smaller because of fierce constraints from the implementation point of view. Thus, we subsequently focus on algorithms attributed to the area of *symmetric cryptography*. Here, the high level of security (112 to 256 bits) and meeting the implementation constraints is possible without requiring unreasonable amounts of keying material. Protocol designers can choose from primitives like block/stream ciphers, hash functions, message authentication codes (MACs), universal hash functions or pseudorandom number generators (PRNGs).

Some of these primitives can be efficiently turned into others. A block cipher can be turned into a hash function. For universal hash functions to be used in the setting for RFID security protocols, a cryptographically secure PRNG is needed to generate new keying material. Standardized and trusted PRNGs are in turn again based on block ciphers or hash functions. This implies that for RFID security protocols using PRNGs on top of primitives like ciphers or hash functions, no substantial additional circuit is

required. MACs are mainly based on hash functions, ciphers or universal hash functions. Summing up this short overview, we conclude that for RFID authentication protocols based on symmetric cryptography either hash functions or ciphers are the most suitable basic building blocks. Subsequently we will thus focus on these.

2.1 Survey of Symmetric Primitives in State-of-the-Art Proposals for RFID Protocols

As mentioned before, the reason to employ cryptographic primitives in RFID protocols is to provide some form of authentication and/or some form of anonymity. In protocols that use hash functions on the RFID tag, the following properties are generally needed. In order to provide some form of anonymity, the output of the hash function should not be distinguishable from a truly random bitstring. For authentication purposes, the designer of RFID protocols relies on the preimage resistance or 2nd-preimage resistance of the used hash function. Occasionally, also collision resistance is needed [13].

In protocols that use ciphers as a cryptographic primitive, anonymity is again provided by the difficulty of distinguishing its output from a truly random bitstring. Authenticity is guaranteed by the resistance against key-recovery attacks of the employed cipher. Subsequently we focus on proposals which offer the possibility to authenticate the tag and/or the reader and thus tackle the problem of tag cloning.

Symmetric Encryption. There are only a few protocol designers which base their RFID authentication protocols on symmetric-key encryption primitives. Feldhofer *et al.* [5] use simple challenge-response authentication for unilateral and mutual authentication of RFID tag and reader. The random values r_t and r_r are the challenges from the tag and the reader which are encrypted using the function E_K which is in their case AES. They mention the problem of key distribution and key management when using symmetric authentication methods but do not provide any detailed solution for it.

Hashing. Weis *et al.* [16] proposed the hash-lock scheme and the improved randomized hash-lock scheme. Thereby, the tag is authenticated by the tuple $(r_t, h(ID, r_t))$ where r_t is a random value generated by the tag for tracking prevention and the hash value is generated over ID and r_t . In the protocol of Henrici *et al.* [7] the tag sends the hash value of its ID together with a transaction number to the reader and authenticates the tag to the reader. The reader responds with a random value which is used to refresh the tag identifier on every successful transaction. Lee and Verbauwhede [9] propose a protocol where a hash function $h(K||r_s)$ of a key K and a random value r_s is used for mutual authentication of reader and tag. Additionally, r_s is used for updating the key in the tag. Dimitriou [3] uses in his authentication protocol a message authentication code (MAC) h_k which is based on hash functions. Random values generated by reader and tag are used for mutual authentication and refreshing the key in the tag. Rhee *et al.* [13] suggest a hash-based challenge-response protocol where the secret key in the protocol is the ID. The tag does not need to update the secret key which avoids attacks by interrupting the session. In the protocol of Choi *et al.* [1] tags have a common secret key K and a tag-specific secret S which are used for mutual authentication based on hash

functions. A counter c is incremented in the tag on each access for prevention of replay attacks.

3 Hardware Implementation of SHA-256

One of the main contributions of this paper is the ASIC implementation of a SHA-256 module which fulfills the requirements for RFID tags. SHA-256 is a member of the SHA-2 family of hash functions which is in turn the latest official descendant of the MD4 family. SHA-256 was chosen because it is widely recommended and offers a security level equivalent to AES-128.

For the design of hardware modules in RFID systems the differences between power consumption and energy consumption is important to notice. In contrast to battery-powered devices where energy consumption is the optimization goal, the mean current consumption is the critical concern for passively powered RFID tags. The duration of the operation is of reduced concern. It is important that the power consumption per clock cycle is limited although the total energy consumption of an operation might be larger. This comes due to the limited energy transmission from the RFID reader to the tag during one clock cycle. Here it is often necessary to serialize operations because the concurrent calculation would exceed the available power and the large voltage drop causes a reset in the circuit. Our implementation goal of the SHA-256 was to minimize the mean power consumption while using only small chip area.

3.1 Description of SHA-256

SHA-256 [11] is an iterated cryptographic hash function based on a compression function that operates on an internal state of 256 bits. This internal state is initialized using IVs as specified in [11]. SHA-256 updates the state of eight 32-bit variables A, \dots, H according to the values of 16 32-bit words M_0, \dots, M_{15} of the message. The compression function consists of 64 identical step transformations as presented in Fig. 1. The step transformations employ the bitwise Boolean functions Maj and Ch , and two GF(2)-linear functions $\Sigma_0(x) = ROTR^2(x) \oplus ROTR^{13}(x) \oplus ROTR^{22}(x)$ and $\Sigma_1(x) = ROTR^6(x) \oplus ROTR^{11}(x) \oplus ROTR^{25}(x)$. The i -th step uses a fixed constant K_i which is a distinct 32-bit word for each step and the i -th word W_i of the expanded message. The message expansion works as follows. An input message is split into 512-bit message blocks (after padding). The message expansion takes as input a vector M with 16 words and outputs a vector W with N words. The words of W_i , the expanded vector, are generated from the initial message M according to the following formula:

$$W_i = \begin{cases} M_i & \text{for } 0 \leq i \leq 15 \\ \sigma_1(W_{i-2}) + W_{i-7} + \sigma_0(W_{i-15}) + W_{i-16} & \text{for } 16 \leq i \leq 63 \end{cases}.$$

The functions $\sigma_0(x)$ and $\sigma_1(x)$ are defined as follows: $\sigma_0(x) = ROTR^7(x) \oplus ROTR^{18}(x) \oplus SHR^3(x)$ and $\sigma_1(x) = ROTR^{17}(x) \oplus ROTR^{19}(x) \oplus SHR^{10}(x)$. After 64 steps, the feed-forward operation is applied. It is done by word by word modular addition of the previous chaining values (the IVs in the case of the first block) to the current state variables.

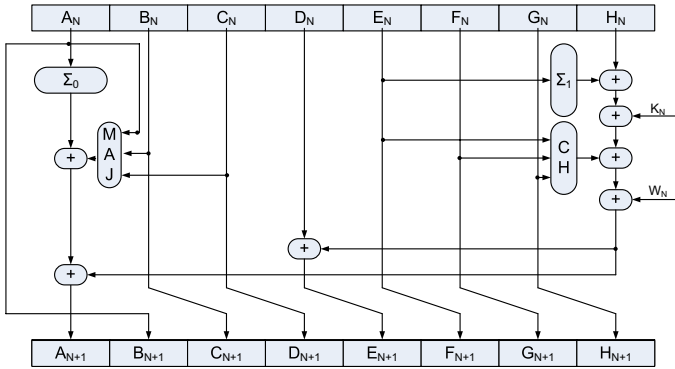


Fig. 1. One step of the state update transformation of SHA-256

3.2 Related SHA-256 Implementations

Although many hash hardware architectures have been proposed in the last years, none of the published work focus on low die-size and low power-consumption requirements as needed for contactlessly powered devices like RFID tags. Nearly all of these architectures focus on GBit throughput rates and do not mind high power consumption at all. Especially the implementations using FPGAs as target technology make extensive use of pipelining and unrolling techniques. Some representatives of this category are Pramstaller and Aigner [12] and Sklavos and Koufopavlou [15]. Only a few publications of ASIC hash implementations are available so far. The implementations of Satho and Tadanobu [14], Dominikus [4] and Dadda *et al.* [2] are directly comparable with our design as they have shown results of SHA-256 implementations.

3.3 Architecture of SHA-256 Module

Implementing the SHA-256 (and also other MD4 family hash functions like SHA-1, MD5 and MD4) algorithm as a 32-bit architecture is the only useful data bit width because of the design of the algorithm. High-level simulations with a data word size of 8 bits showed that the performance is unacceptably bad which was not astonishing as the algorithms were designed for 32-bit platforms. The architecture of the SHA-256 module consists of a datapath and a controller circuit. The datapath of the proposed 32-bit SHA-256 module is depicted in Fig. 2. It is to the best of our knowledge the smallest hardware implementation of the SHA-256 algorithm published so far. The main parts of the module are the RAM circuits, the dedicated logic functions for the SHA-256 transformations, temporary storage registers and one 32-bit adder. The controller module which is not shown in the figure is implemented as a finite state machine that generates the control and address signals for the datapath. All RAM parts have a register-based implementation which allows the use of clock gating to minimize power consumption. The major design principle is that only 32 flip-flops are clocked within the same clock cycle. This averages the power consumption of the circuit which is

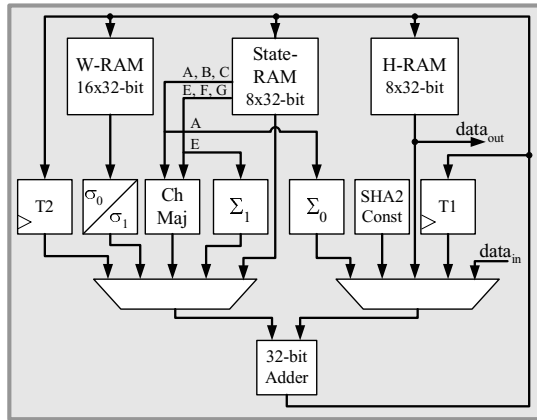


Fig. 2. Architecture of low-power SHA-256 datapath

crucial for wirelessly powered devices like RFID tags. This represents the most important design difference to existing SHA-256 implementations where in every clock cycle all registers of a RAM module need a clock pulse because of the pipelined structure. The RAM consists of the three different parts: message expansion RAM (W-RAM), state variables A-H (State-RAM) and the chaining variables (H-RAM). The W-RAM stores the sixteen 32-bit words W_i necessary for the message schedule. This RAM module is single ported to ease silicon implementation and reduce controlling complexity. The State-RAM contains the eight state variables A-H which are used during the step transformation. Because of the internal structure of the algorithm it was necessary to implement this 8×32 -bit RAM having a separated read and write port. The synthesis results show that the additional hardware for this dual-port RAM is negligible but the throughput is augmented significantly. The third RAM part, the H-RAM, stores the 8×32 -bit chaining variables. It is updated only at the beginning and at the end of each step transformation. The output of the datapath comes directly from the H-RAM.

The dedicated hardware modules in the datapath perform the SHA-256 functions σ_0 , σ_1 , Ch , Maj , Σ_0 and Σ_1 . The inputs of these functions come either from the output of the RAM or they are directly routed from the state variables A-H to the corresponding module. The sequence of 64 constant 32-bit words are stored in a look-up table which was generated by the synthesizer. The two 32-bit registers $T1$ and $T2$ are used during step transformation to store intermediate results. Again clock gating is used to reduce the power consumption while the registers are not required. All RAM circuits, the dedicated hardware modules, the look-up table and the registers have the mechanism to disable the output of the module which sets the 32-bit output of the module to zero. This method, called sleep-mode logic, reduces the switching activity of the combinational logic behind this gates to a minimum. Additionally, the two large multiplexers degrade to merely an OR tree. A further big difference to existing hash architectures is the use of a single 32-bit adder. The critical path of the circuit is heavily reduced to produce less signal activity and hence has lower power consumption.

A calculation of a hash value works as follows. Before the step transformation starts the initial hash value has to be loaded into the H-RAM module. Then the data input including padding is stored in the W-RAM. After transferring data from the H-RAM to the state variables the step transformation starts. According to the SHA-256 algorithm the state variables are updated using the dedicated functions, the intermediate storage registers and the appropriate SHA-256 constant from the look-up table. The message schedule is also executed in each step which allows to generate the required message expansion in place without any additional memory. At the end of the 64 rounds the hash value is calculated using the state variables and the old chaining variables. Now the hash value can be read at the data output or the next message block has to be processed with the same procedure except the initialization of the H-RAM which holds its value.

3.4 Achieved Results

The implementation of our SHA-256 architecture requires a current consumption of $15.87 \mu\text{A}$ at a frequency of 100 kHz on a $0.35 \mu\text{m}$ CMOS process technology with a supply voltage of 3.3 V. All presented results in Table 1 come from simulations on transistor level using Nanosim from Synopsys. Performing a hash calculation on a 512-bit block of data requires 1,128 clock cycles which is suitable for RFID communication protocols using interleaved protocols according to [5]. Although not designed for high speed, the circuit has a maximum clock frequency of 50 MHz and achieves a data throughput of up to 22.5 Mbps. The required hardware complexity is $597,740 \mu\text{m}^2$ which corresponds to 10,868 gate equivalents (GEs). The complexity of each component is also listed in Table 1.

Table 2 presents a comparison of our approach with the work of Dadda [2], Dominikus [4] and Satoh [14]. While Dominikus [4] does not provide information about the implementation of the RAM circuit Dadda [2] and Satoh [14] use shift registers for implementation of the RAM. The state variables, the chaining variables and the expanded message words are in each case structured as a 32-bit pipelined register file where in each clock cycle all registers are updated and cause therefore high power consumption. Hence, power saving techniques like clock gating are not applicable.

It can be seen that the use of very recent process technologies brings advantages in the maximum clock frequency. As RFID tags are low-end devices, today most manufacturer use the cheaper $0.35 \mu\text{m}$ and $0.25 \mu\text{m}$ CMOS process technologies. Unfortunately, all other related work does not include power consumption figures to their results.

4 Discussion of Implications

Table 3 presents a comparison of four hash function implementations with the AES implementation of Feldhofer [6]. The figures of the SHA-256 and SHA-1 implementations are synthesis results while for MD5 and MD4 only estimations are available. Note that the algorithms have different block lengths and security level which was not normalized in the table. The chip area includes all necessary components to operate the module stand alone while the cycle count is the number of clock cycles for one block of data. Kaps *et al.* [8] state that a SHA-1 implementation with only 4,276 gate equivalents which is more energy efficient than their AES implementation is possible. In our

Table 1. Synthesis results of all components of the low-power SHA-256 module

Module/Component	I_{mean} [μA @ 100kHz]	Chip area [GE]
State-RAM	4.14	1,984
W-RAM	3.16	3,881
H-RAM	0.43	2,427
SHA-256 Constants	0.18	612
32-bit Registers	2x0.80	2x197
σ_0 & σ_1 & Σ_0 & Σ_1 & Ch & Maj	1.16	817
32-bit Adder	2.74	156
Multiplexer & others	1.36	233
Controller	1.10	364
Total	15.87	10,868

Table 2. SHA-256 comparison with related work based on power consumption and gate count

SHA-256 ASIC	CMOS Tech. μm	I_{mean} μA @ 100kHz	Chip area GE	f_{max} MHz	Clock cycles
This work	0.35 @ 3.3V	15.87	10,868	50	1,128
Dadda [2]	0.18 @ 1.8V	— ^a	14,000	819	65
Dominikus [4]	0.6 @ 3.3V	— ^a	10,900 + RAM	59	392
Satoh [14]	0.13 @ 1.5V	— ^a	11,484	154	72

^a No figures given. Because of the hardware architecture the current consumption is expected to be a multiple of ours.

setting this analysis does not apply since they do not take the message expansion RAM into account (which compares to 2,400 GEs). They state that an external memory holds these sixteen 32-bit words which is not possible on an RFID tag.

The major outcome is that there are *two* dominating factors that decide on the suitability of symmetric primitives for RFID tags. Firstly, the required number of registers which store state variables, chaining variables and message words. Secondly, the underlying word size of the used primitive. All popular hash functions operate on input blocks of at least 512 bits. Including the size of the state which is at least as big as the output of the hash functions, the resulting lower bound on the gate count is already higher than the smallest known AES implementation.

Since AES operates on elements in $GF(2^8)$, the minimal number of registers (8) that have to be clocked is much lower compared to any member of the MD4 family of hash functions which have a word size of at least 32 bits. Architectures for the MD4 family which operate on less than 32 bits at a time face heavy penalties when taking runtime into consideration. This results in a big difference in the achievable mean power consumption of resulting RFID tags, which is the most critical factor for application of passive RFID tags. Overall, future hash functions designs which aim for efficient implementations on RFID tags need to take both points into account. Consequently, the use of block ciphers like the AES as a basic building block for the design of RFID

Table 3. Synthesis results (for SHA-256, AES, SHA-1) and estimations (MD5, MD4) of hash functions and AES algorithm implementations for low die-size and low power consumption

Algorithm	Chip area GE	I_{mean} μA @100kHz	Clock cycles
SHA-256	10,868	15.87	1,128
SHA-1	8,120	10.68	1,274
MD5	8,400	–	612
MD4	7,350	–	456
AES [6]	3,400	8.15	1,032

security protocols is the most sensible choice as of today. Stream ciphers for resource-constraint environments which are currently evaluated by the project eStream of the European Union sponsored Network of Excellence ECRYPT might be useful as well.

5 Conclusions and Future Work

Contrasting some commonly made assumption in the literature on RFID protocols, we showed the following. Current standards and state-of-the-art low-power implementation techniques favor the use of block ciphers like the AES instead of hash functions as the cryptographic building blocks for secure RFID protocols.

To investigate this issue, we proposed a low-power architecture for the MD4 family of hash functions and gave a detailed analysis of this smallest SHA-256 ASIC implementation known so far. Our result has about 10,000 gates and a mean power consumption of 15 μA at 100 kHz using a cheap 0.35 μm process technology which is indeed *within* the constraints of EPC class 2 tags.

Even though it is interesting to know that also SHA-256 might be small enough for some RFID tags, we have to compare it with implementations of other primitives. Comparing it with the smallest known AES implementation we conclude that no known hash function can achieve similar results given fierce implementation constraints (like mean power consumption) as it is the case for RFID systems. Reducing the gate count by switching to smaller hash functions like SHA-1, MD5 or even MD4 does not affect these conclusions.

The reason is two-fold: Firstly, the underlying word size. The minimal number of registers that have to be clocked in the case of AES (8) is much lower compared to any member of the MD4 family of hash functions which have a word size of 32 bits. This results in a big difference in the achievable mean power consumption of resulting RFID tags, which is an important measure. Secondly, the necessary number of registers. It turns out that all hash functions used today have a structural disadvantage compared to the AES. That is why choosing AES results in comparatively cost-efficient RFID security protocols while keeping a high level of security. Thus it remains to be seen which of the many hash-based protocols can be converted into protocols using a cipher to allow for production of small and cheap tags. Alternatively, AES-based hash functions need to be evaluated from the RFID perspective.

References

1. E. Y. Choi, S.-M. Lee, and D. H. Lee. Efficient RFID Authentication Protocol for Ubiquitous Computing Environment. In *Embedded and Ubiquitous Computing - EUC 2005 Workshops*, volume 3823 of *LNCS*, pages 945–954. Springer, December 2005.
2. L. Dadda, M. Macchetti, and J. Owen. The Design of a High Speed ASIC Unit for the Hash Function SHA-256 (384, 512). In *2004 Design, Automation and Test in Europe Conference and Exposition (DATE 2004)*, volume 3, pages 70–75. IEEE Computer Society, 2004.
3. T. Dimitriou. A Lightweight RFID Protocol to protect against Traceability and Cloning attacks. In *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SecureComm 2005)*, pages 59–66, Athens, Greece, September 2005. IEEE Computer Society.
4. S. Dominkus. A hardware implementation of MD4-family hash algorithms. In *9th IEEE International Conference on Electronics, Circuits and Systems*, volume 3, pages 1143–1146. IEEE Computer Society, October 2002.
5. M. Feldhofer, S. Dominikus, and J. Wolkerstorfer. Strong Authentication for RFID Systems using the AES Algorithm. In *Cryptographic Hardware and Embedded Systems – CHES 2004*, volume 3156 of *LNCS*, pages 357–370. Springer, August 2004.
6. M. Feldhofer, J. Wolkerstorfer, and V. Rijmen. AES Implementation on a Grain of Sand. *IEE Proceedings on Information Security*, 152(1):13–20, 2005.
7. D. Henrici and P. Müller. Hash-based Enhancement of Location Privacy for Radio-Frequency Identification Devices using Varying Identifiers. In *2nd IEEE Conference on Pervasive Computing and Communications Workshops (PerCom 2004 Workshops)*, pages 149–153. IEEE Computer Society, March 2004.
8. J.-P. Kaps and B. Sunar. Energy comparison of AES and SHA-1 for ubiquitous computing. In *2nd IFIP International Symposium on Network Centric Ubiquitous Systems (NCUS 2006)*, *LNCS*. Springer, 2006.
9. Y. Lee and I. Verbauwhede. Secure and Low-cost RFID Authentication Protocols. In *2nd IEEE Workshop on Adaptive Wireless Networks (AWiN)*, 2005.
10. National Institute of Standards and Technology (NIST). FIPS-197: Advanced Encryption Standard, November 2001. Available online at <http://www.itl.nist.gov/fipspubs/>.
11. National Institute of Standards and Technology (NIST). FIPS-180-2: Secure Hash Standard, August 2002. Available online at <http://www.itl.nist.gov/fipspubs/>.
12. N. Pramstaller and M. Aigner. A Universal and Efficient SHA-256 Implementation for FPGAs. In *Austrochip 2004*, pages 89–93, 2004. ISBN 3-200-00211-5.
13. K. Rhee, J. Kwak, S. Kim, and D. Won. Challenge-Response Based RFID Authentication Protocol for Distributed Database Environment. In *Security in Pervasive Computing, Second International Conference, SPC 2005*, volume 3450 of *LNCS*, pages 70–84. Springer, 2005.
14. A. Satoh and T. Inoue. ASIC-Hardware-Focused Comparison for Hash Functions MD5, RIPEMD-160, and SHS. In *International Symposium on Information Technology: Coding and Computing (ITCC 2005)*, volume 1, pages 532–537. IEEE Computer Society, 2005.
15. N. Sklavos and O. Koufopavlou. Implementation of the SHA-2 Hash Family Standard Using FPGAs. *The Journal of Supercomputing*, 31(3):227–248, March 2005.
16. S. A. Weis, S. E. Sarma, R. L. Rivest, and D. W. Engels. Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems. In *1st Annual Conference on Security in Pervasive Computing*, volume 2802 of *LNCS*, pages 201–212, 2003.

An Efficient ID-Based Delegation Network

Taek-Young Youn^{1,*}, Young-Ho Park^{2,**}, Chang Han Kim³, and Jongin Lim¹

¹ Graduate School of Information Security, Korea University, Seoul, Korea
{taekyoung, jilim}@cist.korea.ac.kr

² Dept. of Information Security, Sejong Cyber University, Seoul, Korea
youngho@cybersejong.ac.kr

³ Dept. of Information Security, Semyung University, Jecheon, Korea
chkim@semyung.ac.kr

Abstract. Delegation of signing capability is a common practice in various applications. Mambo *et al.* proposed a proxy signatures as a solution for delegation of signing capability. Proxy signatures allow a designated proxy signer to sign on behalf of an original signer. After the concept of proxy signature scheme was proposed, many variants are proposed to support more general delegation setting. To capture all possible delegation structures, the concept of delegation network was introduced by Aura, and an ID-based delegation network was proposed by Chow *et al.* In the computational point of view, their solution requires E pairing operations and N elliptic curve scalar multiplications where E and N are the number of edges and nodes in a delegation structure, respectively. In this paper, we proposed an efficient ID-based delegation network which requires only E pairing operations. Moreover, we modify our delegation network, and the modified scheme requires only N pairing operations. In general, E is larger than N and E increases according to the complexity of the network. So the modified delegation network is remarkably efficient than the previous delegation network.

Keywords: ID-based Signatures, Proxy Signatures, Delegation Network.

1 Introduction

There are many applications where a delegation of signing capability is required. For example, a traveling executive can delegate to his secretary to sign certain documents during his absence. When a document is given to the secretary, he signs on the document on behalf of the executive. Besides, managers can delegate to their subordinates to perform certain signatures for some insignificant work. An authorized subordinate can sign on a document if the work is in the range of delegated signing capability. In this case, managers can reduce the amount of work [8]. These examples show that delegation of signing capability can be

* This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

** Corresponding author.

useful in various organization. A solution for the delegation of signing capability is the proxy signatures introduced by Mambo *et al* in [8]. Originally, it is devised to delegate the original signer's signing capability to a single authorized person. However, such a simple structure of delegation is insufficient for some applications. We can consider the following scenarios: (1) When an original signer delegates his signing capability to a single proxy, the proxy signer could misuse his delegated ability. In this case, the original signer may want to distribute the signing capability. (2) We can consider the case where the original signer is not a single person but a set of members. For example, a group of members choose a member as a representative of the group, and give their power of attorney to the representative (3) Since a hierarchical structure is common in organizations nowadays, some kinds of chained delegation is required for some application. For example, a president of a company may delegate his signing capability to a vice-president, and the vice-president also delegates his delegated signing capability to a head of department. In a large enterprise, there are many documents to be signed, and some of documents are not significant to be signed by the president. So the signing capability can be hierarchically decentralized. In this case, multi-hierarchical delegation structure is required.

Many kinds of proxy signatures, such as multi-proxy signature [13,9], threshold proxy signature [10,5,12,11,6], proxy multisignatures [7] and multi-proxy multisignature [4] schemes, are proposed to solve the above scenarios. Each scenario can be solved by existing schemes, but the previous schemes are not fully general and flexible to solve all of the scenarios simultaneously. The most generalized and flexible model is the delegation network proposed by Tuomas Aura [3], and an ID-based delegation network was proposed in [2]. The scheme in [2] is fully general and flexible ID-based delegation network and so the scheme can cover all delegation structures.

1.1 Contributions

In this paper, we proposed an efficient ID-based delegation network. The delegation network in [2] requires E pairing operations to certifying E edges, and N elliptic curve scalar multiplications to certifying N nodes. Our proposed delegation network requires only E pairing operations, i.e., we reduce the computational cost by N elliptic curve scalar multiplication. Moreover, by modifying our scheme, we design a delegation network that requires only N pairing operations. Since N is the number of users in a delegation structure, the computation complexity of our delegation network does not increased according to the complexity of the network as the previous delegation network. So the modified scheme is remarkably efficient than the previous ID-based delegation network.

2 Preliminaries

2.1 Bilinear Maps

Let $(\mathbb{G}_1, +)$ and (\mathbb{G}_2, \cdot) be two cyclic groups of prime order q . The bilinear pairing is given as $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$, which satisfy the following conditions:

Bilinearity: For all $P, Q, R \in \mathbb{G}_1$, $e\langle P + Q, R \rangle = e\langle P, R \rangle e\langle Q, R \rangle$, and $e\langle P, Q + R \rangle = e\langle P, Q \rangle e\langle P, R \rangle$.

Non-degeneracy: There exists $P, Q \in \mathbb{G}_1$ such that $e\langle P, Q \rangle \neq 1$.

Computability: There is an efficient algorithm to compute $e\langle P, Q \rangle$ for all $P, Q \in \mathbb{G}_1$.

Such a bilinear map is called an admissible bilinear map. Let P be a generator of \mathbb{G}_1 , and $a, b, c \in GF(q)$. We are interested in the following problems:

Definition 1. (*BDHP: Bilinear Diffie-Hellman problem*) Given a generator P of a group \mathbb{G}_1 and a 2-tuple (aP, bP) , the BDHP is to compute abP .

Definition 2. (*D-BDHP: Decisional Bilinear Diffie-Hellman problem*) Given a generator P of a group \mathbb{G}_1 and a 4-tuple (P, aP, bP, cP) , the D-BDHP is to decide whether $c = ab \pmod q$.

2.2 Security Model for Our Delegation Network

Informally, the security of our delegation network is equivalent to the BDHP in the random oracle model. In this paper, we prove the security by construct an algorithm, which solves BDHP, by using the ability of a delegation network forger \mathcal{F} . To use the ability of \mathcal{F} , we should simulate the attack environment of \mathcal{F} . So, in the proof, we should response the following queries from the forger:

Hash Queries on Oracle H_1 for Identity: When a forgery asks for the public key of user ID_i , we response the query by returning a point $Q_i \in \mathbb{G}_1$ that satisfying $H_1(ID_i) = Q_i$.

Hash Queries on Oracle H_2 for Message: When a forgery asks for the hashed value of a message m , we response the query by returning a point $M \in \mathbb{G}_1$ that satisfying $H_2(m) = M$.

Private Key Extraction Queries: When a forgery asks the private key for some user ID_i , we response the query by returning the secret key d_i of user ID_i where $d_i = sQ_i$. In this case, we assume that the forgery also asks the hash query on oracle H_1 for identity ID_i .

Proxy Signing Key Generation Queries: When a forgery asks the proxy signing key for some user ID_i , we response the query by returning the proxy signing key SK_i of user ID_i .

Final Signing Queries: When a forgery asks the final signature for a delegation structure, we response the query by returning a signature which corresponds to the delegation structure.

Proxy Key Extraction Queries and Final Signing Queries can be seen as a sequence of signing queries. So, we can substitute a sign query for two queries.

Signing Queries: When a forgery asks the signature of ID_i for some message m_j , we response the query by returning the corresponding signature σ satisfying the verification step. In this case, we assume that the forgery also asks the hash

query on oracle H_1 and H_2 , since we should compute Q_i and M_j to generate the signature σ .

For the security of our delegation network, we assume the same model of adversary considered in [2], which models the extremely strong case where the adversary attacks against a single honest user with private keys of all other users. Formal definition of the delegation network forger \mathcal{F} is defined as follows:

Definition 3. *A delegation network forger \mathcal{F} is defined as a forger who produces a signature $(\sigma, ID_1, \dots, ID_N, m_1, \dots, m_E)$ for a delegation structure satisfying the following conditions. In this case, we assume that the network structure is composed of N nodes and E edges where the nodes mean the users in the network and the edges mean the delegation state between two users.*

1. *The signature pass the verifications step.*
2. *There exists $i \in \{1, \dots, N\}$ in which the forger does not ask private key extraction queries for ID_i*
3. *The pair $(\sigma, ID_1, \dots, ID_N, m_1, \dots, m_E)$ has not been presented to the final signing queries. Especially, the pair (ID_i, m_j) is not asked as a signing query.*

Intuitively, a delegation network is secure against the delegation network forger only if the possibility that the forger can generate a set of valid delegation signature is negligible. As we comment, we assume the extremely strong adversary who attacks a single honest user with private keys of all other users. Formal definition of the security of a delegation network against the attacker, explained in Definition 3, is as follows:

Definition 4. *Let \mathcal{F} be a $(q_{H_1}, q_{H_2}, q_S, q_E)$ -forger for a delegation network where q_{H_1} , q_{H_2} , q_S and q_E are the hash query for identity, hash query for message, signing query and private key extraction query, respectively. In this case, the advantage of \mathcal{F} , $Adv_{\mathcal{F}}^{dnf}$, is defined as the probability that the forger \mathcal{F} produces a signature for a delegation network structure which satisfying the conditions stated in definition 3. Then the delegation network is secure against the delegation network forger \mathcal{F} if and only if there exists a negligible function $negl(\cdot)$ such that for sufficiently large k , $Adv_{\mathcal{F}}^{dnf}(q_{H_1}, q_{H_2}, q_S, q_E) < negl(k)$.*

3 Proposed Delegation Network

In this section, we propose an ID-based delegation chain, and then we propose an ID-based delegation network on it.

3.1 Construction of Delegation Chain

Let U_0 be an original signer, U_i be a proxy signer, and ID_i be an identity of U_i . Let V be a verifier.

Setup. In this step, PKG generates groups \mathbb{G}_1 and \mathbb{G}_2 of order prime q , chooses a generator $P \in \mathbb{G}_1$ and a random $s \in GF(q)$, and computes $Q = sP$. He chooses two cryptographic hash functions $H_1 : \{0, 1\}^* \rightarrow \mathbb{G}_1$ and $H_2 : \{0, 1\}^* \times \mathbb{G}_1 \rightarrow \mathbb{G}_1$,

and an admissible bilinear map $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$. Then the public system parameters are $(\mathbb{G}_1, \mathbb{G}_2, P, Q, H_1, H_2, e)$, and the corresponding secret key is s .

Key Generation. U_i requests to PKG for his private signing key. Then PKG computes $d_i = sQ_i$ where $Q_i = H_1(ID_i)$ and sends it to U_i in a secure way. U_i can verify the key by checking the following condition: $e\langle d_i, P \rangle = e\langle Q_i, Q \rangle$. If the condition holds, U_i accepts d_i as his secret key.

Hierarchical Proxy Key Generation. Initially, we let SK_0 as d_0 , and set W_0 as a null string.

1. U_{i-1} chooses a random r_{i-1} and computes $R_{i-1} = r_{i-1}P$.
2. U_{i-1} computes $W_i = H_2(W_{i-1} || ID_{i-1} || ID_i || m_{i-1,i})$ where $m_{i-1,i}$ is the warrant that states the delegation condition from U_{i-1} to U_i .
3. U_{i-1} gives $S_i = SK_{i-1} + r_{i-1}W_i$, \mathbb{M}_{i-1} and \mathbb{R}_{i-1} to U_i where $\mathbb{M}_{i-1} = \{m_{0,1}, \dots, m_{i-1,i}\}$ and $\mathbb{R}_{i-1} = \{R_0, \dots, R_{i-1}\}$.
4. U_i verifies S_i by checking $e\langle P, S_i \rangle = e\langle Q, \sum_{k=0}^{i-1} Q_k \rangle \prod_{k=0}^{i-1} e\langle R_k, W_{k+1} \rangle$. If the above conditions holds, U_i accepts S_i as a part of his proxy signing key and computes his proxy signing key as $SK_i = S_i + d_i$.

Final Signature Generation (Proxy Signing by U_n). For given message m , U_n computes $M = H_2(W_n || ID_n || m)$, and $r_n M$. Then he sends the signature $(\sigma, \mathbb{R}_n, \mathbb{M}_n)$ to V where $\sigma = r_n M + SK_n$.

Final Signature Verification. V computes $A = e\langle R_n, M \rangle$, $B = e\langle Q, \sum_{k=0}^n Q_k \rangle$ and $C = \prod_{k=0}^{n-1} e\langle R_k, W_{k+1} \rangle$. V accepts the signature if the following equations holds: $e\langle P, \sigma \rangle = ABC$.

3.2 Delegation Network for Proxy Signature

Note that, our delegation network can be seen as the extension of the delegation chain. From now, we use the following symbols for convenient: $E_{A,B}$ is an edge that connect U_A with U_B , AE_A is the set of all ancestor edges of U_A , PE_A is the set of all parent edges of U_A , AN_A is the set of all ancestor nodes of U_A , PN_A is the set of all parent nodes of U_A , $\mathbb{M}_A = \{m_{i,j} | (i, j) \in AE_A\}$ and $\mathbb{R}_A = \{R_{i,j} | (i, j) \in AE_A\}$.

Setup and Key Generation: Identical with the description in Section 3.1.

Hierarchical Proxy Key Generation (Delegation from A to B)

Initially, we let W_O is the null string and $SK_O = d_O$ where $d_O = sQ_O$ is the secret key of an original signer U_O .

1. U_A chooses a random $r_{A,B}$ and computes $R_{A,B} = r_{A,B}P$.
2. U_A computes $W_{A,B} = H_2(\sum_{i \in PN_A} W_{i,A} || ID_A || ID_B || m_{A,B})$, and $r_{A,B}W_{A,B}$ where $m_{A,B}$ is the warrant that states the delegation condition from U_A to U_B .
3. U_A gives $(S_{A,B}, \mathbb{M}_A, m_{A,B}, \mathbb{R}_A, R_{A,B})$ to U_B where $S_{A,B} = SK_A + r_{A,B}W_{A,B}$.
4. U_B verifies $S_{A,B}$ by checking the following condition:

$$e\langle P, S_{A,B} \rangle = e\langle Q, \sum_{i \in AN_B} (n_i Q_i) + Q_A \rangle \prod_{(i,j) \in AE_B} e\langle R_{i,j}, W_{i,j} \rangle,$$

where n_i is the number of j which satisfying $E_{i,j} \in PE_A$, and $PE = PE_A \cup E_{A,B}$. If the above conditions holds, U_B accepts $S_{A,B}$ as a part of his proxy signing key and so the proxy signing key is $(r_{A,B}, SK_B)$ where $SK_B = S_{A,B} + d_B$.

Final Signature Generation (by U_R)

1. U_R chooses a random r_R and computes $R_R = r_R P$.
2. U_R computes $M = H_2(\sum_{i \in PN_R} W_{i,R} || ID_R || m)$, and $\sigma = SK_R + r_R M$, for given message $m \in \{0, 1\}^*$.
3. U_R sends the signature $(\sigma, \mathbb{M}_R, m, \mathbb{R}_R, R_R)$ to V .

Final Signature Verification

V accepts the signature if the following equations holds:

$$e\langle P, \sigma \rangle = e\langle Q, \sum_{i \in AN_R} (n_i Q_i) + Q_R \rangle e\langle R_R, M \rangle \prod_{(i,j) \in AE_R} e\langle R_{i,j}, W_{i,j} \rangle,$$

where n_i is the number of j which satisfying $E_{i,j} \in PE_R$.

3.3 Security of the Proposed Delegation Network

Theorem 1. *Let \mathcal{F} be a forger that $(\epsilon, q_{H_1}, q_{H_2}, q_E, q_S)$ -breaks the delegation network where q_{H_1} , q_{H_2} , q_E and q_S are the number of H_1 -hash query, H_2 -hash query, private key extraction query and signing query, respectively. Then, there exists an algorithm \mathcal{A} that solves BDHP with probability ϵ' where $\epsilon' \approx \epsilon \frac{1}{e} \sqrt{\frac{1}{e} \frac{1}{q_S}}$.*

Proof. Suppose that there is a delegation network forger \mathcal{F} who can forge a valid signature. Then we can construct an algorithm \mathcal{A} that solves the computational BDHP by using \mathcal{F} . Let (P, aP, bP) be a given challenge for computational BDHP problem. \mathcal{A} runs the forger \mathcal{F} and simulates its attack environment. \mathcal{A} sets (P, aP) as the public key of the delegation network. To avoid collision and consistently responds to queries, \mathcal{A} maintains two lists H_1 and H_2 . The list is initially empty. \mathcal{A} responds to \mathcal{F} 's queries as follows:

Hash Queries on Oracle H_1 for Identity: For this query, \mathcal{A} gives identical responses to identical queries, maintaining lists relating to its previous hash query responses for consistency. \mathcal{A} responds to \mathcal{F} 's query on ID_i as follows:

1. If ID_i was previous H_1 -query, \mathcal{A} recovers $(ID_i, k_i, Q_i, c_{1,i})$ from its H_1 -list.
2. Else, \mathcal{A} generates a random $c_{1,i} \in \{0, 1\}$ so that $Pr[c_{1,i} = 0] = \delta_1$ for δ_1 to be determined later. If $c_{1,i} = 0$, \mathcal{A} generates a random k_i and set $Q_i = k_i P$; else, it generates a random k_i and set $Q_i = k_i (bP)$.
3. \mathcal{A} adds the tuple $(ID_i, k_i, Q_i, c_{1,i})$ to the H_1 -list and responds to \mathcal{F} as $H_1(ID_i) = Q_i$.

Hash Queries on Oracle H_2 for Message: \mathcal{A} gives identical responses to identical queries, and maintaining H_2 -list. \mathcal{A} responds to \mathcal{F} 's query on a tuple (W_i, m_i) as follows. Note that, m_i is a message that includes the information of sender's identity, so we can find out the intended signer for the queried message tuple (W_i, m_i) . We assume that the intended signer as ID_j .

1. When ID_j is not a previous H_i -hash query, \mathcal{A} treats ID_j as a H_1 -hash query.
2. If (W_i, m_i) was previous query, \mathcal{A} recovers $(W_i, m_i, l_i, M_i, c_{2,i})$ from H_2 -list.
3. Else if $coin_{1,j} = 1$, \mathcal{A} generates a random $c_{2,i} \in \{0, 1\}$ so that $Pr[c_{2,i} = 0] = \delta_2$ for δ_2 to be determined later. When $c_{2,i} = 0$, \mathcal{A} generates a random l_i and set $Q_i = l_i(bP)$;
4. Otherwise, it generates a random l_i and set $Q_i = l_iP$.
5. \mathcal{A} adds (W_i, m_i, l_i, M_i) to the H_2 -list and responds to \mathcal{F} as $H_2(W_i, m_i) = M_i$.

Private Key Extraction Queries: When a forgery asks the private key for some user ID_i , \mathcal{A} sees the H_1 -list, and recovers $(ID_i, k_i, Q_i, c_{1,i})$. If $c_{1,i} = 0$, \mathcal{A} computes the secret key of ID_i as $d_i = aQ_i = a(k_iP) = k_i(aP)$, else aborts.

Signing Queries: When \mathcal{F} asks for signature of user ID_i on a message tuple (W_j, m_j) , \mathcal{A} sees the H_1 -list and H_2 -list, and recovers $(ID_i, k_i, Q_i, c_{1,i})$ and $(W_j, m_j, l_j, M_j, c_{2,j})$. Then, \mathcal{A} responds to the signing queries as follows:

1. If $c_{1,i} = 0$, \mathcal{A} chooses a random r and set the corresponding signature as (σ, R, M_j) where $\sigma = k_i(aP) + rM_j$, and responds to \mathcal{F} as (σ, R, M_j) .
2. Else if $c_{1,i} = 1$ and $c_{2,i} = 0$, \mathcal{A} chooses a random r and computes $\sigma = rl_j(bP)$ and $R = rP - l_j^{-1}k_i(aP)$, and responds to \mathcal{F} as (σ, R, M_j) .
3. Otherwise, \mathcal{A} aborts the simulation.

From now, we measure the probability that \mathcal{A} succeeds to simulate the forger \mathcal{F} , and so it solves the BDHP. The probability that \mathcal{A} responds to all private key extraction queries and to all signing queries are $\delta_1^{q_E}$ and $(1 - (1 - \delta_1)(1 - \delta_2))^{q_S}$, respectively. So the probability that \mathcal{A} succeeds in the simulation is $\delta_1^{q_E}(1 - (1 - \delta_1)(1 - \delta_2))^{q_S}$. To solve BDHP, we expect \mathcal{F} generates a valid delegation signature that includes a signature of ID_i , which is not queried in a private key extraction queries, for a message tuple (W_j, m_j) where $coin_{2,j} = 1$. It is assumed that \mathcal{F} forges a valid signature with probability ϵ and the signature includes a signature of ID_i , which is not queried in a private key extraction queries. The probability that the forged signature includes a signature of ID_i for a tuple (W_j, m_j) with $coin_{2,j} = 1$ is $\frac{1}{2^{k'}}(1 - \delta_2)$ where k' is a security parameter. The term $\frac{1}{2^{k'}}$ comes from the probability that \mathcal{F} will fail to guess $H_2(W_j, m_j)$ without querying H_2 and $1 - \delta_2$ is the probability the corresponding random coin $coin_{2,j}$ is 1. So, we can use the forged signature to solve the BDHP with probability $\epsilon \frac{1}{2^{k'}}(1 - \delta_2)$. Consequently, \mathcal{A} solves the BDHP with probability $\epsilon' = \epsilon \delta_1^{q_E}(1 - (1 - \delta_1)(1 - \delta_2))^{q_S}(1 - \delta_2)$. Let $\delta_1 = 1 - \frac{1}{q_E}$ and $\delta_2 = 1 - \frac{1}{q_S}$. Then ϵ' can be written as $\epsilon' = \epsilon(1 - \frac{1}{q_E})^{q_E}(1 - \frac{1}{q_E q_S})^{q_S} \frac{1}{q_S}$. Note that $(1 - \frac{1}{x})^x \approx \frac{1}{e}$ for large x . Then we can estimate the probability ϵ' as follows:

$$\epsilon' = \epsilon(1 - \frac{1}{q_E})^{q_E}(1 - \frac{1}{q_E q_S})^{q_S} \frac{1}{q_S} = \epsilon(1 - \frac{1}{q_E})^{q_E} \sqrt[q_E]{(1 - \frac{1}{q_E q_S})^{q_E q_S}} \frac{1}{q_S} \approx \epsilon \frac{1}{e} \sqrt[q_E]{\frac{1}{e} \frac{1}{q_S}}$$

We assume that \mathcal{F} makes a forged delegation signature σ on a network structure. For convenient, we assume the secret key of ID_{i^*} is not queried in a private key extraction queries and the corresponding message (W_{j^*}, m_{j^*}) is queried in a hash query on oracle H_2 with $coin_{2,i^*} = 1$. In the following equation, we set $AN_R^* = AN_R/\{i^*\}$. Since the signature passes the verification, we can see the following equation holds:

$$\begin{aligned}
 e\langle P, \sigma \rangle &= e\langle P, SK_R + r_R M \rangle \\
 &= e\langle Q, \sum_{i \in AN_R} (n_i Q_i) + Q_R \rangle e\langle R_R, M \rangle \prod_{(i,j) \in AE_R} e\langle R_{i,j}, W_{i,j} \rangle \\
 &= e\langle P, a(\sum_{i \in AN_R} (n_i k_i) + k_i^* b + k_R) P \rangle e\langle l_R R_R + \sum_{(i,j) \in AE_R} (l_{i,j} R_{i,j}), P \rangle \\
 &= e\langle P, a(\sum_{i \in AN_R} (n_i k_i) + k_i^* b + k_R) P + l_R R_R + \sum_{(i,j) \in AE_R} (l_{i,j} R_{i,j}) \rangle.
 \end{aligned}$$

Then we can see that the σ can be written as follows:

$$\sigma = a(\sum_{i \in AN_R} (n_i k_i) + k_i^* b + k_R) P + l_R R_R + \sum_{(i,j) \in AE_R} (l_{i,j} R_{i,j})$$

Since we know all the k, l and n , we can find abP as follows:

$$abP = k_i^{*-1} (\sigma - (\sum_{i \in AN_R} (n_i k_i) + k_R) (aP) - l_R R_R + \sum_{(i,j) \in AE_R} (l_{i,j} R_{i,j})). \quad \square$$

3.4 A Small Variant of the Proposed Delegation Network

In [2], the authors reserve an open problem to minimize the storage or bandwidth requirement for delegation scenario, by using an efficient aggregate signatures scheme such as the scheme in [1]. If we set $(P, r_A P)$ as the public key of a user U_A and r_A as the corresponding secret key rather uses a random secret value, it does not required to transmit \mathbb{R} with a signature. Since \mathbb{M} should be included in the signature, removing \mathbb{R} from the signature for a delegation network is optimal to minimize the storage and bandwidth. However, this solution requires N more steps to verifying the public key of N users. Since the verification of public key is very costly process, the solution is not desirable.

By adopt the setting of [1], we can enhance the efficiency of the delegation network, we proposed in section 3.2. If we fix the random value chosen by a user in the proxy signing key generation step, we can reduce the size of \mathbb{R} from E to N points of elliptic curve. Though each user uses a fixed random value, anyone can believe that a user participate in a generation of delegation signature, since each user also uses the secret key sQ , where $Q = H_1(ID_A)$ for some user U_A . However, if each user uses a fixed random value, the signature scheme can not guarantee the robustness against a kind of replay attack, i.e, an adversary can send the same signature several times and the signatures are accepted as valid. In a delegation network, same message does not signed, since the message includes the delegation condition, and the delegation period is also included in the delegation condition. So, for a delegation network, we can use a deterministic signatures. The security of the modification of our delegation network, which uses a fixed random secret value, can be proved similar to Theorem 1 with some technique used in [1]. So, it is reasonable to accept that the use of fixed secret value is secure. When each user uses a fixed random secret value, the size of \mathbb{R} is N rather than E . As we commented, the size of E is increased according to the complexity of the network structure. Hence $E - N$ is large for complicate network structure. Moreover, we can reduce the computational cost. The number of pairing operations is also reduced, since we can treat the signatures of U_i at once. So the computation cost is reduced from E to N pairing operations.

3.5 Efficiency

The comparison of our scheme with the existing delegation network [2] is summarized in Table 1. We denote a pairing operation and an elliptic curve scalar

multiplication by P and M , respectively. Let Pt and Wt be the bit size of a point of elliptic curve and a warrant, respectively. We denote E_A and N_A the number of ancestor edges and nodes of U_A in a delegation structure, respectively. Similarly, E_F and N_F are the number of all edges and nodes in a delegation structure, respectively. From now, we denote *Delegation Network 1* and *Dele-*

Table 1. Comparison of our delegation networks with the existing delegation in [2]

	Delegated Key Verification - U_A	Proxy Sign Key Generation - U_A	Final Signature Verification	Length of Signature
Previous Scheme [2]	$(E_A + 2)P + N_A M$	3M	$(E_F + 2)P + N_F M$	$(E_F + 1)Pt + N_F Wt$
Delegation Network 1	$(E_A + 2)P$	2M	$(E_F + 2)P$	$(E_F + 1)Pt + N_F Wt$
Delegation Network 2	$(N_A + 2)P$	1M	$(N_F + 2)P$	$(N_F + 1)Pt + N_F Wt$

gation Network 2 as *DN1* and *DN2* in short. As seen in the table, *DN1* is more efficient than that of [2]. The essential difference between the previous result and our delegation network is the number of scalar multiplications. Our delegation network does not require any elliptic curve scalar multiplications in delegated key verification and final signature verification steps, and our scheme requires only 2M for proxy sign key generation rather than 3M. The length of signature is the same. *DN2* is more efficient than *DN1*. As seen in Table 1, the number of pairing operations for the previous delegation network [2] and *DN1* depend on the number of ancestor edges of a user U_A or all edges in a delegation network structure. For example, in the previous delegation network and *DN1*, U_A should compute $(E_A + 2)$ pairing operations in the delegated key verification step. However, in *DN2*, U_A computes only $(N_A + 2)$ pairing operations. Since, in general $N_A < E_A$, *DN2* is more efficient than the others. In this case, the size of signature is also reduced by $(E_F - N_F)Wt$.

4 Conclusion

In this paper, we proposed an ID-based delegation network. Our proposed scheme requires E pairing operations rather than E pairing operations and N scalar multiplication. Moreover, we design a modified delegation network that requires only N pairing operations. In general, E is larger than N and E increases according to the complexity of the network. So the modified delegation network is remarkably efficient than the previous delegation network.

References

1. Dan Boneh, Craig Gentry, Ben Lynn, and Hovav Shacham, *Aggregate and Verifiably Encrypted Signatures from Bilinear Maps*, EUROCRYPT 2003, LNCS 2656, pp.416-432, Springer.
2. Sherman S.M. Chow, Richard W.C. Lui, Lucas C.K. Hui, and S.M. Yiu, *Identity Based Delegation Network*, Mycrypt 2005, LNCS 3715, pp.99-115, 2005.

3. Tuomas Aura, *On the Structure of Delegation Networks*, PCSFW 1998, IEEE Computer Society Press, 1998.
4. Shin-Jia Hwang, and Chiu-Chin Chen, *New Multi-Proxy Multi-Signature Schemes*, Applied Mathematics and Computation, 147(1):57-67, 2004.
5. Min-Shiang Hwang, Eric Jui-Lin Lu, and Iuon-Chang Lin, *A Practical (t, n) Threshold Proxy Signature Scheme Based on the RSA Cryptosystem*, IEEE Transactions on Knowledge and Data Engineering, VOL.15, NO.6, November/December 2003.
6. Wen-Chung Kuo, and Ming-Yang Chen, *A Modified (t, n) Threshold Proxy Signature Scheme Based on the RSA Cryptosystem*, Proceedings of the Third International Conference on Information Technology and Applications (ICITA05), 2005.
7. Xiangxue Li, Kefei Chen, Longjun Zhang, and Shiquan Li, *Proxy Structured Multisignature Scheme from Bilinear Pairings*, ISPA 2004, LNCS 3358, pp. 705-714, 2004.
8. M. Mambo, K. Usuda, and E. Okamoto, *Proxy Signatures for Delegating Signing Operation*, Proc. 3rd ACM Conference on Computer and Communications Security, 1996.
9. So-Young Park, and Sang-Ho Lee, *Multi-proxy Signatures Based on Diffie-Hellman Problems Allowing Repeated Delegations*, HSI 2005, LNCS 3597, pp. 340-344, 2005.
10. H.-M.Sun, N.-Y.Lee, and T.Hwang, *Threshold proxy signatures*, IEE Proc. Computers and Digital Techniques, Vol.146, No.5, September, 1999.
11. Guilin Wang, Feng Bao, Jianying Zhou, and Robert H. Deng, *Comments on "A Practical (t, n) Threshold Proxy Signature Scheme Based on the RSA Cryptosystem"*, IEEE Transactions on Knowledge and Data Engineering, VOL. 16, NO. 10, OCTOBER 2004.
12. Qingshui Xue, and Zhenfu Cao, *A Threshold Proxy Signature Scheme Using Self-Certified Public Keys*, ISPA 2004, LNCS 3358, pp.715-724, 2004.
13. Q. Xue and Z. Cao, *Improved of Multi-proxy Signature Scheme*, Proceeding of International Symposium on Communications and Information Technologies, pp.450-455, 2004.

Enabling Practical IPsec Authentication for the Internet

Pedro J. Muñoz Merino, Alberto García-Martínez, Mario Muñoz Organero,
and Carlos Delgado Kloos

Universidad Carlos III de Madrid, Department of Telematics Engineering,
Avda de la Universidad, 30
E-28911 Leganés (Madrid), Spain
{pedmume, alberto, munozm, cdk}@it.uc3m.es

Abstract. There is a strong consensus about the need for IPsec, although its use is not widespread for end-to-end communications. One of the main reasons for this is the difficulty for authenticating two end-hosts that do not share a secret or do not rely on a common Certification Authority. In this paper we propose a modification to IKE to use reverse DNS and DNSSEC (named DNSSEC-to-IKE) to provide end-to-end authentication to Internet hosts that do not share any secret, without requiring the deployment of a new infrastructure. We perform a comparative analysis in terms of requirements, provided security and performance with state-of-the-art IKE authentication methods and with a recent proposal for IPv6 based on CGA. We conclude that DNSSEC-to-IKE enables the use of IPsec in a broad range of scenarios in which it was not applicable, at the price of offering slightly less security and incurring in higher performance costs.

1 Introduction

The aim of IPsec ([1], [2], [3], [4]) is the provision of confidentiality, integrity and authenticity. IPsec defines two new types of headers at the IP level: AH (Authentication Header) that provides authenticity and integrity, and ESP (Encapsulating Security Payload) that provides data confidentiality. IPsec can be used in transport mode, in which all the transport and upper layer information is protected, but not the IP header, or tunnel mode, in which the whole IP packet is protected. Although VPNs make intensive use of tunnel mode IPsec between routers, IPsec has not reached widely adoption for protecting end-to-end communications. Furthermore, IPsec implementations like Opportunistic Encryption method proposed by the open source project FreeS/WAN [5] have not been widely adopted.

There are many reasons that explain the low acceptance of IPsec in end-to-end communications [6]. One major obstacle for the adoption of IPsec is the authentication mechanisms available. The traditional methods for authentication are Pre-Shared Key and Digital Signatures (for the rest of the paper the terms Digital Signatures or Certificates will be used to refer to the same authentication mechanism). The use of Pre-Shared Key is only feasible when the number of communicating hosts is small, because of the difficulties of distributing of the key to each pair of hosts. Therefore, this solution is not valid for the casual communications that occurs in the Internet. On

the other hand, the use of certificates requires a common Certificate Authority (CA) between two hosts, but at present no unique CA hierarchy has been adopted in the Internet. Therefore, this solution is not valid between two hosts that do not trust in a common CA. However, new possibilities for authentication have arisen such as the ones derived from the proper use of the Cryptographically Generated Addresses (CGA, [7]), although in this case IPv6 is required.

As it can be inferred from the previous paragraph, present authentication methods for IPsec can only be applied to hosts that conform to some restrictive conditions. In this paper, we try to solve the limited scope of present authentication methods for IPsec. We propose a new authentication method for IKE, DNSSEC-to-IKE that allows authentication for IPsec in a global basis through Internet. This is because the DNSSEC-to-IKE method only require from the hosts to have access to the DNSSEC security infrastructure, provided that a hierarchical DNSSEC infrastructure over Internet exists. With the inclusion of this authentication method one barrier to the usage of IPsec in casual communications could be removed.

DNSSEC-to-IKE is based on the use of DNSSEC security architecture and the ubiquitous reverse DNS infrastructure to store securely the public key that corresponds to a given IP address. The few changes required to IKE to support this mechanism guarantee easy deployment.

We perform a comparison between DNSSEC-to-IKE and state-of-the-art authentication mechanisms, to determine the conditions in which each mechanism can be applied, the security provided and the performance of the validation. We conclude that DNSSEC-to-IKE provides convenient authentication for casual communication between Internet hosts without requiring specific infrastructure or costly key distribution.

The remainder of this paper is organized as follows. In Section 2 there is an outline of the general framework for IPsec authentication; both the state-of-the-art authentication methods and the CGA authentication mechanism are described. Section 3 explains the DNSSEC-to-IKE authentication method. Section 4 makes a comparative analysis of the authentication methods in terms of applicability, security, and performance of the validation process. Finally, Section 5 is devoted to the conclusions.

2 General Overview of the IPsec Authentication Framework

IPsec needs to establish a session security context in order to be used in AH and ESP headers. Although any set of protocols can be used for this functionality, the ones recommended for IPsec are: ISAKMP [8], [9], IKE [10] and OAKLEY [11].

In this work, the main mode was selected for IKE Phase 1. The main mode requires the six ISAKMP messages for each of the three authentication methods. The first two messages negotiate security parameters and the authentication interchange. The next two messages generate a shared Diffie-Hellman key. Finally, the last two messages authenticate the Diffie-Hellman key previously exchanged. Results can be easily extended for aggressive mode due to the similarity of the authentication process.

Among the three authentication methods described by IKE, we do not analyse in detail the Public Encryption method because it results in the same scalability

problems as the Pre-Shared Key one, since it is necessary for each host to know the public key of the rest of the hosts to which it communicates. In the following subsections the authentication model of the IKE protocol is briefly explained for Pre-Shared Key, Digital Signatures [10], and CGA.

2.1 IKE Authentication with Pre-shared Key

In the fifth and the sixth messages of the IKE exchange, the hosts can verify the identity of each other by checking the hashes received that are bound to the Pre-Shared Key agreed for the communication between these two hosts.

2.2 IKE Authentication with Certificates

When Digital Signatures (Certificates) are used, previously to the fifth message, each host can request a certificate to the correspondent host - the certificate is not needed if the host knows in advance the public key of the other host. In the request, a host indicates to the correspondent a list with the CAs in which it trusts. Next, in messages number 5 and 6, the hosts exchange an *Authorization Payload*, which is a signature with its private key. The messages also include a *CERT Payload* with certificates belonging to the CAs contained in the request. The host must check the authenticity of the certificates in order to validate the public key of the other host, and next the Authorization Payload is validated with the public key obtained from the certificate.

2.3 IKE Authentication with CGA

CGA, defined in [7], are IPv6 addresses that incorporate into the 64-bit interface identifier a cryptographic one-way hash of a public key and a prefix owned by the node, creating a binding between this public key and the resulting address. An enhancement to IKE to allow CGA-based authentication is described in [12] as a modification to the Digital Signatures method. In this case, the CERT payload contains the public key and all the parameters required to reconstruct the interface identifier of the address, and therefore validate the authenticity of the IPv6 address of the sending host. The Authorization Payload is signed with the private key associated to the public key of the CGA. Therefore, a host can be authenticated as the legitimate owner of an IPv6 address by checking the validity of the CGA and verifying the Authorization Payload signature with the public key received.

3 DNSSEC-to-IKE Authentication Method

We propose a new method for authentication through IKE, named DNSSEC-to-IKE. This new method can provide authentication to scenarios for which previous authentication schemes were not appropriate. We first present a brief overview of DNSSEC, from which the DNSSEC-to-IKE method derives its authentication infrastructure. Then we describe the DNSSEC-to-IKE method in detail, including the modifications required for the IKE exchange.

3.1 DNSSEC Overview

DNSSEC [13], [14], [15] defines a set of new registers for authenticating the information that is stored in the DNS. The added registers are four:

- DNSKEY, a public key that is associated to a DNS zone.
- RRSIG, a signature of any register of a specific DNS zone with its private key associated to the public key available in DNSKEY. For the rest of the paper we denote the signature of a register R as SIG(R).
- NSEC, which proves that some information does not exist in the DNS.
- DS, a hash of the DNSKEY register of a subzone.

Additionally, an IPSECKEY register for storing IPsec keying material associated to the name to be resolved in DNS [16] has also been added.

An example of how a host can obtain a register associated with a name when DNSSEC is used is the following: The querier host issues a request for a given Fully Qualified Domain Name to the root zone, for which it must know the $DNSKEY_0$ public key. The resolver sends the query to a root server, and receives a response containing the $DNSKEY_0$, $SIG_0(DNSKEY_0)$, NS_0 – Name Server –, $SIG_0(NS_0)$, DS_0 and $SIG_0(DS_0)$ registers. All of these signatures use the $DNSKEY_0$ of the root zone, as it is referred by the notation SIG_0 , so the host, that previously knows the public key for this zone, can validate all the received registers ($DNSKEY_0$, NS_0 and DS_0). The NS_0 register allows the querier host to access to the server in charge of the subzone requested. Also, the DS_0 register provides a hash of the $DNSKEY_1$ of the first subzone. In general, the resolver will communicate with the server referred in the NS_i register to obtain its $DNSKEY_{i+1}$ and the information of the following subzone $i+1$. The hash of this $DNSKEY_{i+1}$ is compared with the DS_i register obtained from the higher level i DNSSEC server. By this way, the $DNSKEY_{i+1}$ public key of each subzone can be authenticated. The previous steps are repeated as many times as intermediate DNSSEC servers are between the root and the final local server of the host to be resolved.

3.2 Description of the DNSSEC-to-IKE Authentication Method

In this paper, we detail a method for conveying in IKE authentication information based on the use of the reverse DNS infrastructure along with the DNSSEC facilities to provide the public key corresponding to the IP address of a given host for which an IPSECKEY register has been configured, as required by IPsec authentication. The authentication information that is exchanged through IKE includes the chain of successive zone delegations until the leaf zone that corresponds to the IP address in the reverse DNS is reached, along with the signatures stored in the DNSSEC. Then a host can authenticate the IP-based identity for a corresponding host requiring only basic DNSSEC parameters such as the public key of the root zone. The correspondent node will use the root zone public key to validate the chain of zone keys until the public key of the other host is validated. Note that this validation process does not necessarily force accesses to the DNS, since all the information required can be conveyed in the IKE protocol.

The security provided by the DNSSEC-to-IKE method has to be analysed in administrative terms. It is important to note that IPsec authentication requires the

guarantee that a given host is the legitimate owner of a given IP address. Therefore, if an infrastructure is built to perform this IP identity proof, it should be assured that the infrastructure grants the rights to claim for the IP identity to the legitimate owner. The reverse DNS assigns a zone to each IP address by reversing the address in IP notation, and prepending the resulting string to the in-addr.arpa or ip6.arpa suffix, defining a correspondence between an IP address and a reverse DNS leaf zone. The RIRs (Regional Internet Registries¹) guarantee that only the administration responsible for a range of addresses is assigned the management of the corresponding reverse DNS zone. Reverse DNS zone management is further delegated to clients as addresses are. The assignment of a stable address to a given host is also the responsibility of the administration in charge of the most specific address range. Then, through the appropriate coordination between the end-host administrator and the end-site administrator, the information stored in this reverse DNS infrastructure should correspond to the legitimate IP user.

Next we describe the details of the modifications of IKE to convey this information. Since a chain of successive DNSSEC authentication registers should be interchanged and a list of acceptable DNSSEC root servers for authentication should be suggested, then this is analogous to the Digital Signature authentication method but certificates and CAs are used instead of DNSSEC registers and DNSSEC root servers. For this reason, the IKE Digital Signature method can be taken as a starting point for the definition of the DNSSEC-to-IKE method. For the meaning of the *Certification Request Payload*, *CERT Payload* and *Authentication Payload* and its use inside the IKE Digital Signatures authentication method, the correspondent RFCs should be consulted [8], [10]. Next, we describe how these fields should be used for the DNSSEC-to-IKE method.

The validation process can be accelerated if the verification of some parts of the certificate chain could be omitted. For example, a host in the same network segment as the correspondent knows securely the public key in the DNSSEC for the segment, so there is no need to receive and validate the whole certificate chain beginning from the root zone. Therefore, a *Certification Request Payload* is used to suggest acceptable roots for the authentication, with the following different cases:

- 1) A Certification Request Payload is sent with a list of zones of the reverse DNS domain name corresponding to the IP address. In this case, the requester knows the DNSKEY public key for any of this zones, so the answer can include the security chain starting from any of this zones (preferably, the chain starting from the most specific zone should be sent to reduce the authentication payload).
- 2) A Certification Request Payload is sent with an empty zone name. In this case, a host is requesting all the authentication information of all the zones the other host holds, preferably starting from the root DNS zone to avoid the need for accessing to the DNS to perform the validation.
- 3) The Certification Request Payload is not sent. This occurs when the sender already has the public key of the other host, for example because it obtained it through DNSSEC.

¹ The RIRs have been delegated the responsibility for managing the assignment of addresses, reverse DNS zones and autonomous system numbers, in their corresponding regions. Currently there are five RIRs: RIPE, ARIN, APNIC, LACNIC and AFRINIC.

For the each previous cases of Certification Request Payload, we have to consider the following CERT Payload formats:

- 1) If a Certification Request Payload is sent with a list of zones, then the CERT will contain only one block of DNSSEC-to-IKE information, with the information related to the zone with a more specific or complete name.
- 2) If the Certification Request Payload is sent with an empty zone name, then the CERT will contain only one block of DNSSEC-to-IKE information with the information correspondent to the zone with a shorter name.
- 3) If a Certification Request Payload is not sent, then the CERT will be empty.

A block of DNSSEC-to-IKE information contains the chain of successive DNSSEC registers required by a host that knows the DNSKEY of the zone of the reverse DNS name to authenticate to finally obtain the IPSECKEY of the host to be authenticated. Such a block is inserted into the CERT Payload and includes the chain of data composed by N pairs of information of this type:

$$\text{DNSKEY}_{i_i}, \text{SIG}_{i_i} (\text{DNSKEY}_{i_i}), \text{DS}_{i_i}, \text{SIG}_{i_i} (\text{DS}_{i_i})$$

N is the number of DNSSEC successive subzones from a root zone to the final zone. All these zones are contained in the name stored in reverse DNS (for example the 4.3.2.1.in-addr.arpa name contains the following subzones: arpa, in-addr.arpa, 1.in-addr.arpa and so on). Moreover, the block includes the DNSSEC information of the public key of the IP address from the final zone:

$$\text{IPSECKEY}_{i_N}, \text{SIG}_{i_N} (\text{IPSECKEY}_{i_N})$$

In the same way, a resolver DNSSEC host that knows the DNSKEY public key of the root zone can obtain the DNSSEC chain of authentication directly from the DNSSEC infrastructure instead of requesting it in the Certification Request Payload of IKE. However, including all the certification information in the IKE messages reduces the latency for host authentication.

The Authentication Payload includes a signature of the data used to authenticate the entity of the ID payload. In the DNSSEC-to-IKE, it is signed with the private key associated to the public key stored in the DNSSEC server (IPSECKEY_{i_N} register).

4 Comparative Analysis

In this section, a comparative analysis is performed between different authentication methods in terms of security, performance, and applicability. The methods considered for the analysis are Pre-Shared Key, Certificates, the IKE extensions for CGA-based authentication, and the DNSSEC-to-IKE method proposed in this paper.

4.1 Requirements

In this subsection, we discuss the requirements for applying each authentication method, so that it can be understood in which scenarios can be applied.

The Pre-Shared Key method requires the secure distribution of the agreed key before the communication.

The requirements for applying the Certificates method to authenticate a host are the following:

- A shared trusted CA in the hierarchy path of both the requesting host and the host to be authenticated
- Proper management for the Certification Authorities ensuring proper password security, revocation lists, etc.

The basic requirement for applying the CGA is that the authenticated address must be an IPv6 address.

The requirements for applying the DNSSEC-to-IKE authentication method are:

- The host to be authenticated must have an associated public key signed by its zone
- The requesting host must have at least the DNSKEY public key of a reverse DNS zone to which the host to be authenticated belongs, being the root zone a special case for this.

It may be required for the requester to implement resolver functions if the host to be authenticated does not include in the CERT payload the whole chain of signatures beginning at the root key, because the requester would need to use the DNS protocol to access to the keys of the upper zones in the hierarchy. However, this can be removed by requiring at the host to be authenticated to configure the whole chain to its reverse DNS name.

It is relevant to highlight that the deployment of an authentication mechanism for IPsec does not require additional infrastructure to the one naturally provided for securing the DNS, except for the configuration of the key of the reverse name corresponding to the end-host.

With these considerations, we can derive some conclusions:

CGA allows authentication between hosts that do not share any common information, and does not require any kind of infrastructure. These two valuable features are not available in the other authentication methods. Unfortunately, the IPv6 restriction limits drastically the number of practical scenarios in which the requirements are fulfilled, although it could be useful in the future if IPv6 is deployed.

Pre-Shared Key is limited to few hosts, so it cannot be considered as a solution for broad Internet support.

The Certificate based method can be applied for all the hosts that trust a common CA. Consequently, its applicability is restricted, since a common CA infrastructure is not currently available for authenticating any pair of hosts in the Internet.

Finally, the DNSSEC-to-IKE method enables the possibility of authenticating any pair of hosts in the Internet, since the DNSSEC is expected to provide an Internet-wide trust infrastructure. Although the DNSSEC infrastructure has not been fully deployed, recent advances have been made for the specific support of reverse DNS zones: the reverse DNS zones depending from RIPE (www.ripe.net) are signed since January 2006. The deployment of DNSSEC for reverse DNS zones by the RIRs would enable access to the reverse DNS with the authentication being provided by a small number of keys (as many as RIRs).

4.2 Security

The specific risks for the authentication method based in Pre-Shared Key depend on the security of the key distribution process, on the hash functions and on the private keys exchanged, apart from the risks associated to the management of the private key in the host. Pre-Shared Key authentication does not depend on the trust of a third party.

The security of the Certificates method is based on several aspects: the trust model used for issuing certificates from a Certification authority to another or to a final user, the security provided by the private keys of the CAs, the private keys of the entities to be authenticated, the revocation lists, and the appropriate management of the private keys in all the CAs and the authenticated host. In this authentication mechanism, the user must trust in one or several third parties.

For CGA-based authentication, the weakest security element is imposed by the limitation to 64 bits in the interface identifier of the resulting address. An attacker can try to generate private/public key pairs until the 64-bit hash containing the public key equals to the legitimate one. CGA provides some means of making more difficult this attack through a SEC parameter contained in the IPv6 address that imposes an additional condition to the CGA structure, requiring the last $16 \cdot \text{SEC}$ bytes of a hash different from the one used to obtain the address to be 0. Then, an attacker willing to hijack the CGA identity requires $O(2^{59+16 \cdot \text{SEC}})$, ranging SEC from 0 to 7. The condition expressed by the SEC parameter affects only the time to generate the CGA, not the time required for validating its identity. Additionally, the security of the public key used for the hash, and proper management of the private key are relevant for the protection provided by this authentication method.

The elements affecting the security of the DNSSEC-to-IKE mechanism are the trust model for delegating reverse DNS domains to the administrations responsible for the corresponding addresses, the strength of the public key of the host and of the DNS zones, and the security of the private keys used by each element in the authentication chain. Additionally, several signatures generated by the private keys are public, so some attacks to the private keys used for signing are enabled.

As a conclusion, the security offered by the presented authentication methods greatly varies. The Pre-Shared Key authentication method is the most robust, since the key used can be as strong as required, and the distribution method can be devised to involve a few entities, therefore reducing risks. Certificates are also robust, allowing proper selection of the key length, although any of the several entities involved could be compromised, leading to a security disruption. DNSSEC-to-IKE is similar to Certificates, but it does not allow revocations, providing lower security. Finally, CGA provide the weakest security, because the interface identifier, that is the main cryptographic token for the authorisation process, is limited to 64 bits, although the security can be increased by proper configuration of the SEC parameter.

4.3 Performance Time

In this section we estimate the computing time that is necessary for authenticating the other host in IKE. For the meaning of the different parameters the IKE RFC [10] should be consulted. D symbolizes decryption and V validation.

The performance time for checking the authenticity of a host with Pre-Shared Key is T and it can be calculated, being prf a pseudorandom function, (in case no prf is negotiated in IKE, then prf corresponds to the negotiated hash in IKE), as follows:

$$\begin{aligned} T &= T1+T2+T3+T4+T5+T6 \\ T1 &= T[\text{SKEYD}] = T[\text{prf}(\text{pre-shared-key}, N_{i_b} \mid N_{r_b})] \\ T2 &= T[\text{SKEYD_d}] = T[\text{prf}(\text{SKEYID}, g^{xy} \mid \text{CKY-I} \mid \text{CKY-R} \mid 0)] \\ T3 &= T[\text{SKEYD_a}] = T[\text{prf}(\text{SKEYID}, \text{SKEYID_d} \mid g^{xy} \mid \text{CKY-I} \mid \text{CKY-R} \mid 1)] \\ T4 &= T[\text{SKEYD_e}] = T[\text{prf}(\text{SKEYID}, \text{SKEYID_a} \mid g^{xy} \mid \text{CKY-I} \mid \text{CKY-R} \mid 2)] \\ T5 &= T[\text{HASH_I}] = T[\text{prf}(\text{SKEYID}, g^{xi} \mid g^{xr} \mid \text{CKY-I} \mid \text{CKY-R} \mid \text{SA}_{i_b} \mid \text{ID}_{ii_b})] \\ T6 &= T[\text{D}(\text{payloads})] \end{aligned}$$

$T1$ is the time to compute SKEYD which is a string derived from secret key material that is only known by the two hosts (this time is calculated in a slightly different way in Pre-Shared-Key and Certificates). $T2$, $T3$ and $T4$ are the times to compute different keys that are used in ISAKMP. $T5$ is the time to compute a specific hash. $T6$ corresponds to the time to decrypt the payloads of messages 5 and 6 of the IKE exchange. $T2$, $T3$, $T4$, $T5$ and $T6$ also appear in the estimation of the authentication cost for the rest of the methods.

The performance time T for checking the authenticity of a host with Digital Signatures (certificates) is calculated as:

$$\begin{aligned} T &= T1+T2+T3+T4+T3+T4+T5+T6+T7+T8 \\ T1 &= T[\text{SKEYID}] = T[\text{prf}(N_{i_b} \mid N_{r_b}, g^{xy})] \\ T7 &= V(\text{SIGN}) + M \\ T8 &= N * T[\text{check one certificate}] = N * (V[\text{certificate}] + \text{hash}[\text{certificate}] + M) \end{aligned}$$

$T2$, $T3$, $T4$, $T5$ and $T6$ have the same meaning as in the Pre-Shared-Key method while $T1$ is calculated a bit different. $T7$ is the time for validating the signed message that is on the Authorization Payload. $T8$ represents the time for checking the certificates of all the CAs that chain from the common trusted root. N is the number of CAs in this path, including the common trusted root. $V[\text{certificate}]$ is the time to validate the encrypted hash of the certificate using the public key of a CA. $\text{hash}[\text{certificate}]$ is the time to compute the hash of the certificate and M to check if it matches with the decrypted before. The cost expressed in $T8$ is also incurred when the host does not receive the certificates from the correspondent host but directly from the CAs.

The performance time for checking the authenticity of a host with CGA can be calculated as:

$$\begin{aligned} T &= T1+T2+T3+T4+T3+T4+T5+T6+T7+T8 \\ T1 &= T[\text{SKEYID}] = T[\text{prf}(N_{i_b} \mid N_{r_b}, g^{xy})] \\ T7 &= V(\text{SIGN}) + M \\ T8 &= \text{hash1}(\text{public key}, \text{parameters}) + \text{hash2}(\text{public key}, \text{parameters}) + M \end{aligned}$$

The first seven expressions are the same as in the Certificate method. The difference for CGA is shown in $T8$, which represents the time required to validate the CGA address checking the association to the public key that has been received. Then, it represents the time for performing the hash1 operation as it is defined in the RFC for CGAs [7], and the hash2 operation if SEC is not 0, and matching the first result with the interface identifier of the address.

The performance time for checking the authenticity of a host with the DNSSEC-to-IKE method is computed as follows:

$$T = T1+T2+T3+T4+T5+T6+T7+T8$$

$$T1 = T[\text{SKEYID}] = T[\text{prf}(\text{Ni}_b \parallel \text{Nr}_b, g^{xy})]$$

$$T7 = V(\text{SIGN}) + M$$

$$T8 = N * T[\text{check registers of one server of the chain}] = N * T[(V[\text{DNSKEY}] + \text{hash}[\text{DNSKEY}] + M + V[\text{DS}] + \text{hash}[\text{DS}] + M)]$$

$$T9 = T[\text{check final registers}] = T[V[\text{IPSECKEY}] + \text{hash}[\text{IPSECKEY}] + M]$$

The first seven expressions are analogous to the method with certificates. T8 represents the time for validating the DNSKEY and DS registers that form the chain of authenticity if N zones are considered. T9 is the time for validating the final IPSECKEY register.

We have measured the CPU time required for the validation with each authentication method in a Pentium 4, 2.1 GHz with Windows XP. For each authentication method, different authentication cases were considered. For each authentication case, the different IKE messages were generated, considering typical payload values. The execution time is the medium time between all the considered cases. For each case, the time is calculated adding the different execution times of each cryptographic operation performed in the PC. It was used for IKE: Diffie-Hellman group 2 (1024 bits), DES as encryption algorithm, HMAC-SHA-1 as hash function and also as prf function. X.509 has been used for certificates and RSA-1024 as digital signature algorithm. The results are shown in Table 1.

Table 1. Time required for the different authentication methods

Authentication method	Execution time (in microseconds)
Pre-Shared-Key	25
CGA	223
Certificates with N=1 (1 CA)	396
Certificates with N=4 (4 CAs)	966
DNSSEC with N=1 (1 intermediate zone)	762
DNSSEC with N=4 (4 intermediate zone)	1870

From the expressions shown above and the times presented in Table 1, we can derive some conclusions. The fastest method is the Pre-Shared Key because only five hash operations and a decryption have to be computed. The second fastest one is CGA, that adds the cost a signature and two hash calculation to the previous case. The most expensive methods are Certificates and DNSSEC-to-IPsec, and the validation time depends on the number of hierarchical elements between a root and the final authority for both CAs and DNSSEC-to-IKE. Being the number of intermediate elements in the hierarchy the same, then certificates will execute about two times faster than DNSSEC because in DNSSEC it is necessary to perform two validation operations per each hierarchical element, while for certificates only one is required.

5 Conclusions and Future Work

In this paper we have presented a new authentication method for IPsec, DNSSEC-to-IKE for the inclusion of DNSSEC based authentication in the IKE exchange. This new method provides authentication for the use of IPsec between two end hosts in many situations that were not possible with previous authentication methods if the appropriate DNSSEC infrastructure exists.

The DNSSEC-to-IKE method is a variation from the IKE Digital Signatures authentication specification. The security of the information in which the DNSSEC-to-IKE authentication relies is provided by applying DNSSEC to the reverse DNS infrastructure. The IKE exchange contains a chain of authenticating elements that rely on the key of a zone known for both parties, being in the worst case the DNS root zone key, or a limited number of keys such as the ones provided by the Regional Internet Registries. It should be noted that the deployment of an authentication mechanism for DNSSEC-to-IKE does not require additional infrastructure to the one naturally provided for securing the reverse DNS.

We have presented a comparative analysis of the DNSSEC-to-IKE method with traditional IPsec authentication methods and the recently proposed CGA-based authentication in terms of applicability, security, and computing cost. DNSSEC-to-IKE is the only method that enables practical authentication for any two previously unrelated hosts in the current IPv4 Internet by taking advantage of the security infrastructure that is being built now for securing the reverse DNS (for example, through the signature of the reverse DNS zones depending from the RIRs). CGA provides similar features for the IPv6 Internet at a very low performance cost, although offering weaker security. In scenarios in which PKI is practical, the Certificate based authentication provides better security support than DNSSEC-to-IKE by means of revocation lists, and also provide better performance for the authentication process. Finally, the Pre-Shared Key method can only be applied to sets containing a limited number of hosts, although it provides very good security and excellent performance.

As future work, the proposed method can be tested in different scenarios and it can be integrated with different IPsec implementations. In addition, an algorithm can be proposed to execute in each host in order to determine for each situation in real time what is the best authentication method to use.

Acknowledgement. This work has been partially supported by the Programa Nacional de Tecnologías de la Información y de las Comunicaciones, MEC-CICYT project MOSAIC-LEARNING TSI2005-08225-C07-01 and 02 and by the IMPROVISA project TSI2005-07384-C03-02.

References

1. Kent, S., Atkinson, R.: Security Architecture for the Internet Protocol. RFC 2401, (1998)
2. Kent, S., Atkinson, R.: IP Authentication Header. RFC 2402, (1998)
3. Kent, S., Atkinson, R.: IP Encapsulating Security Payload (ESP). RFC 2406, (1998)
4. Thayer, R., Doraswamy, N., Glenn, R., IP Security Document Roadmap, RFC 2411 (1998)
5. FreeS/WAN Project, <http://www.freeswan.org/>

6. Ionnadis, J.: Why don't we still have IPsec, dammit. NDSS 2003, (2003)
7. Aura, T.: Cryptographically Generated Addresses (CGA). RFC 3972, (2005)
8. Maughan, D., Schertler, M., Schneider, M., Turner, J.: Internet Security Association and Key Management Protocol (ISAKMP). RFC 2408
9. Piper, D.: The Internet IP Security Domain of Interpretation for ISAKMP. RFC 2407, (1998)
10. Harkins, D., Carrel, D.: The Internet Key Exchange (IKE). RFC 2409, (1998)
11. Orman, H.: The OAKLEY Key Determination Protocol. RFC 2412, (1998)
12. Laganier, J.: Using IKE with IPv6 Cryptographically Generated Address. Internet Draft, (2003)
13. Arends, R., Austein, R., Larson, M., Massey, D., Rose, S.: Protocol Modifications for the DNS Security Extensions, RFC 4035 (2005)
14. Arends, R., Austein, R., Larson, M., Massey, D., Rose, S.: Resource Records for the DNS Security Extensions. RFC 4034, (2005)
15. Arends, R., Austein, R., Larson, M., Massey, D., Rose, S.: DNS Security Introduction and Requirements. RFC 4033, (2005)
16. Richardson, M.: A Method for Storing IPsec Keying Material in DNS. RFC 4025, (2005)

Preamble Encryption Mechanism for Enhanced Privacy in Ethernet Passive Optical Networks

Pedro R.M. Inácio^{1,2}, Marek Hajduczenia^{1,3}, Mário M. Freire²,
Henrique J.A. da Silva³, and Paulo P. Monteiro^{1,4}

¹ Siemens S. A., Research and Development Department, Rua Irmãos Siemens, 1
2720-093 Amadora, Portugal

² IT-Networks and Multimedia Group

Department of Computer Science, University of Beira Interior
Rua Marquês de Ávila e Bolama, P-6201-001 Covilhã, Portugal

³ Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Pólo II
3030-290 Coimbra, Portugal

⁴ Instituto de Telecomunicações – Pólo de Aveiro, Universidade de Aveiro
3810-193 Aveiro, Portugal

{pedro.inacio, marek.hajduczenia, paulo.monteiro}@siemens.com,
mario@di.ubi.pt, hjas@ci.uc.pt

Abstract. Ethernet Optical Passive Networks (EPONs), defined as low cost access networks, combine Ethernet technology with an optical fiber infrastructure to deliver voice, video and data services from a Central Office (CO) to end-users. Since all data in the downstream is broadcasted, it is susceptible to be eavesdropped by a malicious user, which can use it to try Theft of Service (ToS) through masquerading techniques. These threats remain present when encryption is applied to EPON frame payloads. In order to avoid user profile inference through data mining techniques, a method for encryption of the preamble of the data units is proposed in this paper and a short description of its operations is presented. This new encryption mechanism assures that any two EPON frames are always transmitted with different and uncorrelated preambles.

Keywords: Security, Ethernet Passive Optical Networks, EPON, encryption mechanism, EPON frame preamble, privacy.

1 Introduction

Ethernet Optical Passive Networks (EPON), defined as low cost access networks, combine Ethernet link-layer protocol with an optical fibre infrastructure to deliver voice, video and data services from a Central Office (CO) to end users. The device terminating the optical fibre in the CO is normally referred to as Optical Line Terminal (OLT) and, at the other end of the Passive Optical Network (PON), end-users are connected through Optical Network Terminals (ONUs), placed in the house or business premises. Data transmissions originated by the OLT are usually known as *downstream*. The term *upstream* is used for transmissions initiated by a given ONU [1].

In order to maintain the cost of the network deployment as low as possible, Passive Splitter Combiners (PSCs) are placed between the OLT and the ONUs. The purpose of these devices in the upstream transmission is to combine several optical signals coming from individual ONUs into a single, shared, fiber channel; or to split the single optical signal coming from the OLT to all connected ONUs. A PSC, as the name suggests, is a completely passive device that does not perform logical operations or amplifies the signal. For those reasons, the number and position of PSCs units has to be carefully considered prior to network deployment.

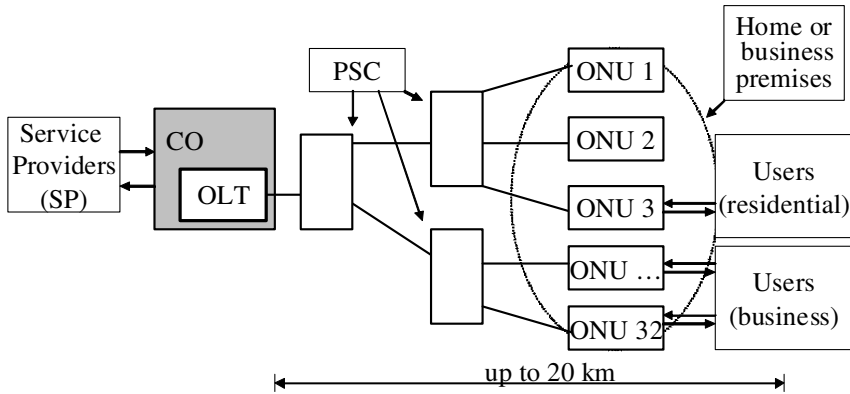


Fig. 1. Tree-and-branch EPON system topology with a number of connected ONUs with various deployment scenarios

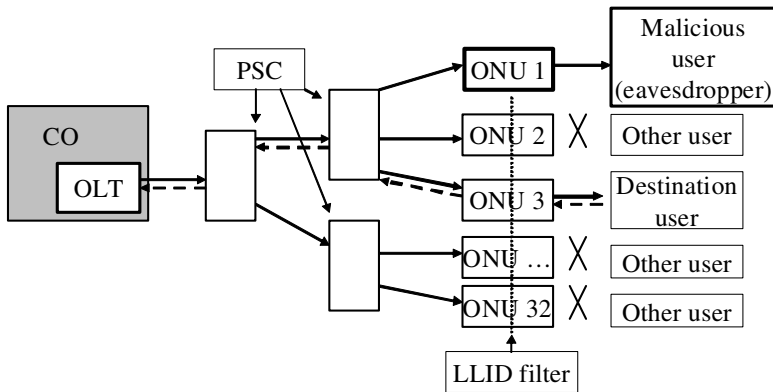


Fig. 2. Upstream (dashed arrows) and downstream (normal arrows) communications in an EPON system. Data originated by the OLT is broadcasted to all ONUs in the system, while data originated by individual ONUs is delivered only to the OLT (unicast link).

Fig. 1 depicts an EPON tree-and-branch deployment, where all the aforementioned elements of the system are logically schematized. The number and localization of ONUs in the figure are merely symbolic, however, depending on several variables,

their number is never superior to 64, neither their distance to the OLT bigger than 20 km.

Due to the physical properties of the PSCs, any signal originating from the OLT will unavoidably arrive to every ONU on the system. Downstream communications are, therefore, of broadcast type (Fig. 2, normal arrows). In the upstream direction, if no reflections occur on the PON system (mainly on the PSCs), the signal originating from an ONU will only be seen by the OLT (Fig. 2, dashed arrows) [1]. As some segments of the optical fiber are shared between more than one ONU, the upstream transmission requires medium access protocol, in the form of the Time Division Multiple Access (TDMA), which in EPON system is supported through application of the Multi Point Control Protocol (MPCP), providing a general framework for TDMA channel access.

Fig. 3 exemplifies how the upstream bandwidth management is performed. MPCP Data Units (MPCP DUs), called GATES [2], emitted by the OLT, inform every ONU about the start time and length of its transmission slots.

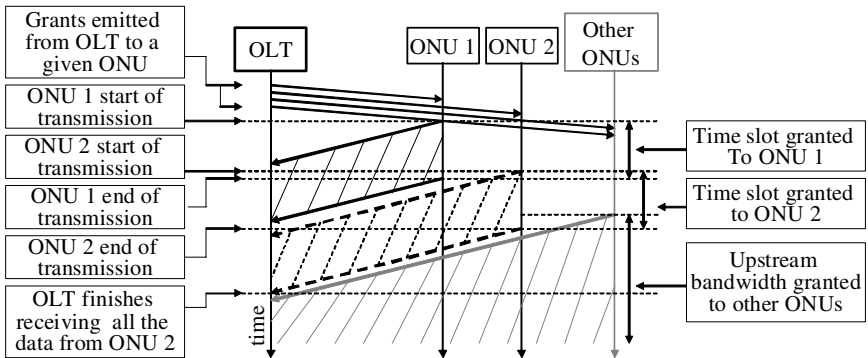


Fig. 3. Upstream bandwidth management using assignment of transmission time slots

The remaining part of this paper is organized as follows. Section 2 describes EPON security threats according to their severity. Section 3 presents the most common security mechanisms used in EPONs. Section 4 proposes a new mechanism preventing data mining and profiling based on the Logical Link Identifier (LLID) values and section 5 presents main conclusions.

2 EPON Security Threats

2.1 High Severity Threats

High severity threats derive directly from the fact that downstream communications are of broadcast type [3], [4]. In such medium, point-by-point connections are only possible through link-layer emulation, where every ONU connected to the system has one or more than one assigned LLID that univocally identifies it in the network. When transmitting, an ONU stamps every frame with one of its LLIDs. Frames sent in the upstream direction carry, consequently, the address of the source of the communication. In the downstream direction, all frames carry the LLID address of the

destination entity, with the exception of the broadcast data units, which are delivered with the so-called *broadcast LLID*.

Fig. 4 emphasizes the differences between Ethernet (upper section of the figure) and EPON frames (lower section of the figure). While on EPONs the preamble of the frame carries addressing information (LLID and its respective Cyclic Redundancy Check 8 (CRC8)), the very same preamble in Ethernet frames is only used to assure proper clock recovery process and data stream alignment. The reserved fields in the EPON frame preamble are not used within the PON context.

When an ONU receives an EPON frame, it applies a filtering policy based on the set of LLIDs assigned to it. If the LLID on the frame coincides with one of the ONU’s LLIDs, the frame is to be accepted and forwarded for further processing; if not, it should be discarded.

A malicious EPON system user, aware of the operation of the LLID filtering rules, can deactivate them on a given ONU, enabling it to work in a promiscuous mode. This procedure will give him the capability to listen to all downstream communications, in a completely unnoticeable manner. Once active, a promiscuous mode allows an attacker to learn Medium Access Control (MAC) and LLID addresses of other ONUs on the EPON system; perform user profiling (quantity and type of traffic) by monitoring LLID, MAC or content information; and infer characteristics about the upstream traffic by observing downstream MPCP DU exchange, especially the GATE MPCP DUs, carrying bandwidth allocation to particular ONUs and LLIDs. Privacy is said to be assured when it is not possible to infer confidential information through passive attacks. For this reason, traffic analysis can be seen as an attempt against privacy [5].

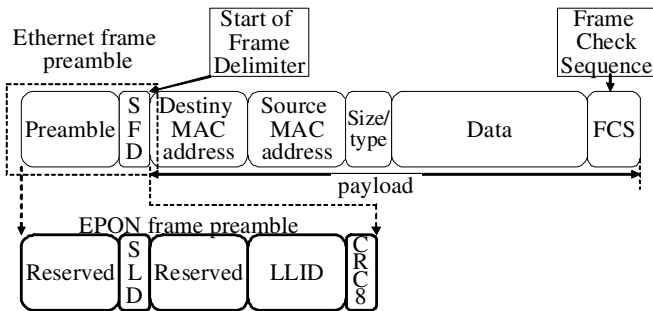


Fig. 4. Ethernet/EPON frame scheme. EPON and Ethernet frames differ only in the preamble format, which in EPON case contains additional Point-to-Point (P2P) emulation data.

2.2 Medium Severity Threats

In the upstream direction, EPON is a Multi-Point-to-Point (MP2P) network. Upstream transmissions are controlled by the OLT through application of the MPCP protocol and of a Bandwidth Allocation Algorithm (BAA). An attacker can masquerade himself as another ONU (by stamping all the emitted frames with an LLID that is assigned to other ONU) and try to theft service from a legal user. He can also obtain access to confidential information or restricted resources or even take advantage of

the operation of the BAA mechanism through spoofed MPCP messages, attempting to reduce the upstream channel bandwidth assigned to other, legitimate users [4].

Another medium severity threat, normally classified as a physical layer threat, is the possibility that an attacker has to superpose the upstream signal from others ONUs with a high power signal, aiming to compromise the optical detection capabilities of the OLT sensor [3], leading the system to a downtime.

2.3 Low Severity Threats

If there are significant reflections on the medium, or if the attacker has technologically advanced devices (with high detection capacities), he can try to eavesdrop frames sent by another ONU [3]. After analyzing them, he can send a modified version of the data units upstream, or simply discard them.

Although theoretically possible, occurrence of optical reflections is highly improbable since their value, caused by the PSCs and fibre channel (more specifically: splices between fibre sections) is extremely low.

3 Currently Available Mechanisms

EPON security has been subject of concern for quite a while [1], [3]-[11], and it is commonly accepted that encryption of the payload of the Ethernet frames solves the problem of confidentiality [3], [4] and of data origin authentication [9]. Symmetric encryption standards, as the Advanced Encryption Standard (AES) or Data Encryption Standard (DES), are commonly used to encrypt private data contents on EPON communications. At the EPON level, content encryption corresponds, in the best case scenario, to the encryption of the Data, FCS, Size/Type and MAC fields. Preamble encryption is not currently performed in any existing EPON implementation [12], [13]. Fig. 5 shows how content encryption can be seen from the data link layer point-of-view.

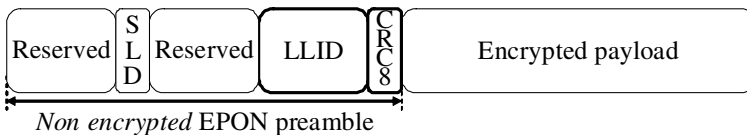


Fig. 5. EPON frame scheme: content encrypted

In the downstream direction, encryption of the Ethernet frames payload prevents a malicious user to access confidential data within the frames. Without the correct decryption keys, using state of the art computers, message reconstruction is virtually impossible; and depending on the size of the encryption keys, their regeneration rate and their secrecy level, the data transported within the EPON data units, can be considered to be more or less secure.

In the upstream direction, the error detection code included in the encrypted payload (a 32 bit wide Cyclic Redundancy Check (CRC32) code transported in the Ethernet frame FCS field) can be used to validate the origin of the data. Decryption of the payload followed by a successful match between the calculated and the decrypted

CRC32 assures, at least, that the party that sent the message had the correct encryption key. If the key, used for encryption and decryption of the content of the frames, was traded using reliable means, and its secrecy is guaranteed, the ONU from which the frame was originated is not faking its identity. In the downstream direction, unless the hypothesis of having a fake OLT is accounted for, it makes no sense to talk about data origin authentication, since in that direction the data can only be originated by the OLT.

The low severity problem of potential upstream reflections has also been addressed by the solution described in [14], which comprises a set of reflecting devices, that should be strategically placed on the PON system to intentionally generate noise. Any unwanted reflection is merged with its non-correlated reflections on the disturbing devices, turning the echoed signal into non decodable.

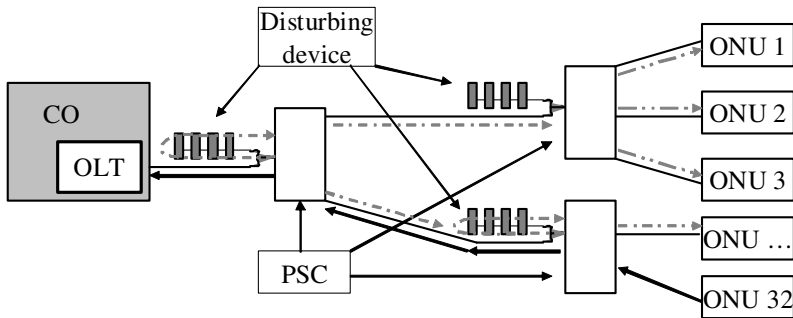


Fig. 6. State of the art solution to protect the EPON system against threats that explore possible upstream reflections

Fig. 6 depicts this physical layer security mechanism. In the example, a fraction of the signal emitted from ONU 32 is intentionally reflected downstream to all ONUs by more than one disturbing device. In the particular case, the signal emitted by ONU 32 is only reflected by the two disturbing devices of the upstream path. The problem of possible upstream reflections can also be solved by encryption of the upstream communications.

On switched networks, destiny information is crucial to correctly forward the data units, from the source to their destiny. Encryption of fields containing such information requires switching devices along the way, to be capable of decrypting the necessary address information before forwarding it. As such process normally decreases the overall networks speed and increases their deployment complexity, letting some of the address information in plain text (destiny address) is typically considered a solution of a choice.

In case of EPONs, the downstream Ethernet frames (even the ones with the contents encrypted) include the LLID information in the preamble of the frames. As the downstream data can be passively observed, a malicious user can still sort the incoming data based on the LLID or CRC8 information. Some implicit/explicit information can be obtained through the observation of such data. For instance, the malicious user will know which are the currently active LLIDs, their activity rate or the downstream bandwidth assigned to them. From the downstream MPCP messages, he can extrapolate the upstream assigned bandwidth for a given LLID. Additionally,

since two different frames with the same LLID are supposedly encrypted with the same key, the collected messages can be also used to feed encryption key searching algorithms or data mining techniques.

4 Anti Data-Monitor Mechanism

As indicated previously, most of the existing security solutions for EPONs are focused on the confidentiality problems. These are, in fact, considered the most severe problems on this kind of networks. However, there is useful information that can be directly deduced from the unencrypted fields within the preamble of the PON frames, or inferred from their analysis. Once a malicious user has gathered and examined sufficient profiling information about an LLID (or from a user associated to it), he can devise better the next step. This constitutes a medium severity problem within the context of EPONs, since private information can be easily obtained by a malicious person.

In this paper, we propose a mechanism for encryption of the EPON frame preambles. This encryption scheme can be applied on EPON systems because the origin and destiny of the information are always known and no switching or routing is performed along the path. The signals are simply separated and sent to all ONUs or combined and sent to the OLT. ID information is only important to the filtering module in the receiving devices and not to data forwarding along the path.

As every downstream frame is submitted to the filtering module in all the ONUs on the system, the proposed mechanism will only adds one additional decryption step before filtering.

In the upstream direction, the encryption mechanism provides data origin authentication at the bottom of the data link layer, providing that the shared keys are unique and their secrecy is assured. In cases where the previous conditions are met, the BAA supports the capability of the OLT to validate the identity of an ONU.

The proposed method offers the best data protection policy to EPON users. If the EPON frames are fully encrypted (preamble and payload), sensitive information is no longer accessible and, consequently, the purpose to eavesdrop the downstream traffic ceases to exist. Users are protected against privacy attempts while the confidentiality of the data is assured by the payload encryption.

Since the computational cost of the solution is to maintain as low as possible, the encryption/decryption functions are based on simple xOR operations.

Table 1. Logical table for xOR operation

\oplus	1	0
1	0	1
0	1	0

$$\begin{aligned} \oplus : (0,1) &\rightarrow (0,1) \\ \oplus (A, B) &= A \oplus B = (A \cap \sim B) \cup (\sim A \cap B) \end{aligned} \quad (1)$$

Table 1 and equation (1) define mathematically the aforementioned function. In the equation, the symbol \sim stands for “negation of”, while \cup represents the bitwise operator *OR* and \cap the bitwise operator *AND*. The *xOR* operator is represented by the \oplus symbol. In the description below, the expression “encryption of sequence A with sequence B” refers to the result of the bit-by-bit application of the bitwise operator *xOR* to the two inputs. Sequence A must be of the same length of sequence B.

4.1 Encryption

Encryption of the sensitive Ethernet frame preamble fields (along with the encryption of the Ethernet frame payload) would significantly reduce the amount of exposed system’s information. In order to exhibit anti-monitor capabilities, the mechanism must meet some criteria. For instance, it has to be suitable for operating at the data link layer; and the cipher texts, resulting from the encryption of two consecutive frames (with the same LLID), have to have an infinitesimal probability of being equal. By other words, once encrypted, two consecutive frames will have always different preambles, even when the respective plain text ones are equal. Once met, this particular property mitigates any type of data mining techniques based on the values of the LLID or CRC8 fields.

EPON frame preamble encryption is depicted in Fig. 7. As the Start of LLID Delimiter (SLD) field does not contain any valuable system information, its encryption is useless and it was not further considered. The method proposed herein assumes that a pair of Secret Keys (SK) (one for the upstream encryption/decryption, one for the downstream encryption/decryption) is shared between each ONU in the PON system and the OLT. The key exchange must be carried out previously, using a secure Key Agreement Protocol (KAP).

When a frame is to be sent from the OLT, or from an ONU, the transmitting party should generate two different keys: a Random Key for the LLID (RK_{LLID}) and a Random Key for the CRC8 (RK_{CRC8}). Assuring that, for each time the keys are generated they are completely random, the bitwise operation *xOR* of each one of them with a static sequence of bits will produce a random sequence as well.

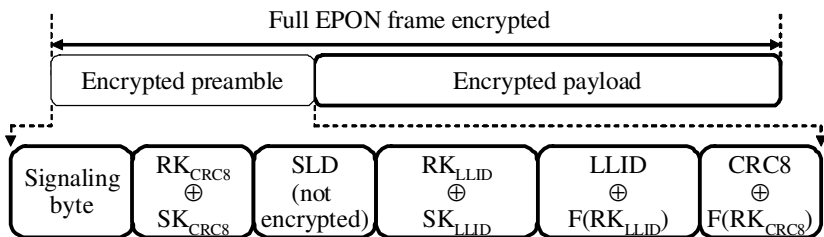


Fig. 7. EPON frame scheme: completely encrypted frame

The LLID field is then encrypted with the bit sequence that results from the application of a non invertible Function (F) to the RK_{LLID} . A reciprocal procedure applies itself for the CRC8 field. In order to prove its worth against data sorting techniques, the F function should produce an output apparently as random as its input. In this case, hash functions fit perfectly the aforementioned requirements.

After encrypted with the SK, the RK_{LLID} and RK_{CRC8} values are conveyed in the non used fields of the preamble (see Fig. 7). The initial byte of the preamble is used to indicate the EPON frame preamble (and even the payload) is encrypted or not (signaling byte in Fig. 7). By these means, the receiving end will know which frames must be decrypted before being filtered. Fig. 7 schematizes a completely encrypted EPON frame where system sensitive fields are directly accessible without previous decryption process, as described in Section 4.2. The encrypted preamble is depicted in detail to emphasize its internal structure.

4.2 Decryption

When a party receives a frame, it should first check the preamble signaling byte. If it indicates that the preamble of the frame is encrypted, the decryption mechanism should be applied to the cipher. An SK should be used to decrypt the RK_{LLID} and RK_{CRC8} , which are fed into the F function. With the resulting sequences, decryption of the LLID and CRC8 fields of the preamble is straightforward.

The OLT must use the SK corresponding to the ONU currently transmitting. The key can be easily retrieved by the BAA that controls the upstream transmissions and, therefore, has the information of what ONU is transmitting at the moment. If, after decryption, the LLID or the CRC8 do not match the expected values, the frame has either an error or it comes from an invalid ONU (perhaps a promiscuous ONU) and should be discarded.

Each ONU uses its respective SK to decrypt downstream encrypted frames. If after decryption, the LLID or CRC8 do not match, the frame was not intended to it and must be discarded.

4.3 Simulation Study

The proposed encryption mechanism was implemented in a distributed, event driven EPON system simulator, which reproduces all the aspects of data transmission in the said networks. All active EPON system components (OLT and ONUs) are represented by individual processes (software programs), operating on individual machines, communicating through sockets and emulating physical level in the EPON system.

The encryption mechanism was implemented on a computer program that simulates an EPON system. The active EPON (OLT and ONUs) components are represented by individual threads that communicate via sockets. Before transmission, the Ethernet frame preamble is encrypted as described herein. At the receiving point, decryption also follows the same specifications: after decryption, the frames are sent to the filtering module which decides about the legitimacy (OLT case) or destiny (ONU case) of the message.

The following lines were taken from one of the sockets log. They are related to the encryption and upstream transmission of EPON frames originated by the same ONU. Each line contains the encrypted EPON frame preamble (written in the hexadecimal notation) and the generated RK used to encrypt it. The plaintext preamble is [55-55-d5-5555-7ffd-2b] and the SK for upstream transmissions for the given ONU is 57-bc91.

```
#1 [05-c8-d5-eb22-d68d-22]  RK:bc-9e59
#2 [05-67-d5-6466-3413-cf]  RK:33-daf6
#3 [05-df-d5-430e-1f73-43]  RK:14-b24e
#4 [05-bf-d5-4bab-56b2-7a]  RK:1c-172e
#5 [05-e2-d5-1d43-1da5-7a]  RK:4a-ff73
#6 [05-79-d5-d664-e7ee-b7]  RK:81-d8e8
#7 [05-e4-d5-043f-2115-6c]  RK:53-8375
#8 [05-2d-d5-8aeb-d610-5b]  RK:dd-57bc
```

It may be concluded that all the preambles of EPON frames originating from the very same LLID are different and statistically uncorrelated. The only fields that are equal in all messages are the SLD, containing the predefined value d5 [2] and the initial field used for signaling purposes and containing, in this case, the value 05. In the simulation, the value 55 is used to indicate that the frame is in the plain text format; a 05 signals preamble encryption; a 50 payload encryption; and 55 stands for complete frame encryption.

Based on the above presented frame preambles it is impossible to say whether if they come from the same LLID or not. This constitutes the perfect example of how the proposed encryption mechanism works. Notice that all the messages were successfully delivered, decrypted and accepted by the filtering module in the OLT.

5 Conclusions

Currently available security mechanisms do not take into account all inherent EPON system threats. The most severe security issues stem from the broadcast type of the downstream communications and, as encryption of the downstream frame payload does not cover the LLID and CRC8 information, it is possible to apply data mining techniques to this data and profile individual network users.

Aiming for the solution of that problem, an anti data-monitoring mechanism was introduced and discussed in this paper. The proposed method not only turns unfeasible any attempt to sort the downstream data, but also assures data origin authentication in the upstream channel (providing that a secure KAP was used to exchange the encryption keys). The cryptographic properties of the algorithm counterbalance its implementation complexity. As the algorithm is very modest, in terms of processor operations and memory requirements, it is suitable for a data link layer implementation, which is the layer on which EPON systems operate.

The decryption mechanism is straightforward and does not impose significant delay on data reception. As defined herein, preamble encryption is faster than payload encryption (simple xOR against AES encryption) and, some of the parts of the proposed encryption procedure can be pre-processed or processed in parallel.

Acknowledgements. The authors would like to thank Fundação para a Ciência e Tecnologia (FCT), Portugal.

References

1. Kramer, G. and Pesavento, G.: Ethernet Passive Optical Network (EPON): Building a Next-Generation Optical Access Network. *IEEE Communications Magazine* (2002) 62-73.
2. IEEE Standard 802.3ah - Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications (2004).
3. Hiironen, O. P. and Pietiläinen, A.: Security Threats and Defense Models in EPON (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/sep02/sec/Pietilainen_sec_1_0902.pdf, Last Access at 15 May, 2006.
4. Pietiläinen, A., Maislos, A., Kramer, G., Hiironen, O. P., and Haran, O.: Security Baseline Proposal (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/hiironen_general_1_0702.pdf Last Access at 15 May, 2006.
5. Hiironen, O. P., Pietiläinen, A., and Nylund, A.: IEEE802.3ah Ethernet in the First Mile, Privacy in EPON (2002). Available at the URL: http://www.ieee802.org/3/efm/public/may02/hiironen_1_0502.pdf, Last Access at 13 June, 2006.
6. IEEE 802.3ah Ethernet in the First Mile Task Force - IEEE 802.3ah P2MP (2002) Presentation Materials. Available at the URL: <http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/index.html>, Last Access at 15 May, 2006.
7. Cook, C., Ford, B., Haran, O., Legoff, Y., Mahalingam, M., Mccammon, K., Michalowski, R., Pietiläinen, A., Fujimoto, Y., Romascanu, D., Sala, D., and Sambasivan, S.: IEEE 802.3ah, Security Objectives for Ethernet Passive Optical Networks (EPON) (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/sep02/sec/FordMccammon_sec_1_0902.pdf, Last Access at 15 May, 2006.
8. Goff, Y. L., Fujimoto, Y., Murakami, K., Haran, O., and Hiironen, O.P.: Encryption layer comparison (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/hiironen_p2mp_4_0702.pdf, Last Access at 15 May, 2006.
9. Hiironen, O. P.: Message Authentication in EPON (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/hiironen_p2mp_2_0702.pdf, Last Access at 13 June, 2006.
10. Kim, J.: Authentication and Privacy in EPON (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/kim_jin_p2mp_3_0702.pdf, Last Access at 15 May, 2006.
11. Murakami, K.: Authentication and Encryption in EPON (2002). Available at the URL: http://grouper.ieee.org/groups/802/3/efm/public/jul02/p2mp/murakami_p2mp_1_0702.pdf, Last Access at 15 May, 2006.
12. PMC-Sierra: PAS5001 EPON OLT for FTTH Broadband Access. Available at the URL: <http://www.pmc-sierra.com/products/details/pas5001/index.html> http://www.pmc-sierra.com/cgi-bin/download_p.pl?res_id=12160&filename=2061016_012120.pdf, Last Access at 15 May, 2006.
13. Teknovus, Inc.: TK3701 Product Brief (2005). Available at the URL: <http://www.teknovus.com/tk3722.html> http://www.teknovus.com/downloads/TK3701_PB.pdf, Last Access at 15 May, 2006.
14. Pohjola, O. P. and Tervonen, A.: Method and System for Secure Upstream Transmissions in Passive Optical Networks (2005). PCT/FI2004/000486.

SMARTCOP – A Smart Card Based Access Control for the Protection of Network Security Components*

Joaquín García-Alfaro¹, Sergio Castillo¹, Jordi Castellà-Roca²,
Guillermo Navarro¹, and Joan Borrell¹

¹ DEIC/UAB, 08193 Bellaterra (Catalonia), Spain
{jgarcia, scastillo, gnavarro, jborrell}@deic.uab.es

² DEiM-ETSE-URV, 43007 Tarragona (Catalonia), Spain
jordi.castella@urv.net

Abstract. The protection of network security components, such as *firewalls* and *Intrusion Detection Systems*, is a serious problem which, if not solved, may lead a remote adversary to compromise the security of other components, and even to obtain the control of the system itself. We are actually working on the development of a kernel based access control method, which intercepts and cancels forbidden system calls potentially launched by a remote attacker. This way, even if the attacker gains administration permissions, she will not achieve her purpose. To solve the administration constraints of our approach, we use a smart card based authentication mechanism for ensuring the administrator's identity. In this paper, we present an enhanced version of our authentication mechanism, based on a public key cryptographic protocol. Through this protocol, our protection module efficiently verifies administrator's actions before granting her the privileges to manipulate a component.

1 Introduction

Network security components, such as *firewalls* and *Intrusion Detection Systems*, are almost always working with special privileges to execute their tasks. This situation can allow remote attackers to acquire these privileges and perform unauthorized activities [2]. The existence of programming errors within the code of these components, the illicit manipulation of their related resources (e.g., processes, executables, and configuration files), or even the increase of privileges through operating system's errors, are just a few examples regarding means in which a remote adversary can bypass traditional security policy controls.

In [4] we presented a protection module integrated into the kernel of an attack prevention system intended to intercept and cancel forbidden system calls launched by a remote attacker. More specifically, the mechanism we presented avoids escalation attacks through an access control scheme which handles the protection of the system's elements. Indeed, this scheme prevents that potentially dangerous system calls (e.g., cancellation of a process) could be produced from one element against another one. The

* This work has been partially funded by the Spanish Ministry of Science and Technology (MCYT) through the projects TIC2003-02041 and SEG2004-04352-C04-01, and the Catalan Government Department DURSI, with its grant 2003FI-126.

protection is hence achieved by incorporating an access control mechanism that may allow or deny a system call based on several criteria – such as the identifier of the process making the call or some of the parameters passed to it.

The approach presented in [4] allows, moreover, to keep away from the necessity of trusting special users with privileged rights, by delegating the authorization for the execution of a given system call to the internal access control mechanisms. Therefore, and contrary to other approaches, it provides a unified solution, avoiding the implementation of different specific mechanisms for each component, and enforcing the compartmentalization principle [10]. This principle is based in the segmentation of a system, so several elements can be protected independently one from another. This ensures that even if one of the elements is compromised, the rest of them can operate in a trusted way. For our job, several elements from each component are executed as processes. By specifying the proper permission based on the process ID, for instance, we can limit the interaction between these elements of the component. If an attacker takes control of a process associated to a given component (through a buffer overflow, for example), she will be limited to make the system call for this given process.

Nevertheless, it is not always possible to achieve a complete independence between the elements. There is a need to determine which system calls may be considered as a threat when launched against an element from the component. This requires a meticulous study of each one of the system calls provided by the kernel of a given operating system, and how they can be misused. We must also define the access control rules for each one of these system calls. For our approach, we proposed the following protection levels to classify the system calls: (1) critical processes protection; (2) communication mechanisms protection; and (3) protection of files associated to the elements.

According to these protection levels, we then presented in [5] a prototype implementation of our kernel based access control mechanism developed for *GNU/Linux* systems and called SMARTCOP (which stands for *Smart Card Enhanced Linux Security Module for Component Protection*). This implementation was developed over the *Linux Security Modules* (LSM) framework [11]. This framework does not consist of a single specific access control mechanism; instead it provides a generic framework, which can accommodate several approaches. It supplies several hooks (i.e., interception points) across the kernel that can be used to implement different access control strategies. Such hooks are: *Task hooks*, *Program Loading Hooks*, *File systems Hooks* and *Network hooks*. This set of LSM hooks can be used to provide protection at the three different levels proposed above.

Furthermore, the LSM framework adds a set of benefits to our implementation. First of all, it introduces a minimum load to the system when comparing it to kernels without LSM, and does not interfere with the normal system activities [11]; second, the access control mechanism can be integrated in the system as a module, without having to recompile the kernel; third, it provides a high degree of flexibility and portability to our implementation when compared to other proposals for the Linux kernel, such as [7] and [9], where the implementation may require some kernel modifications; and fourth, the LSM interface provides an abstraction which allows the modules to mediate between the users and the internal objects from the operating system kernel – to this effect, before accessing the internal object, a hook may call functions provided by a

given module and which may decide whether allow or deny the access to the internal object, for example.

Through the use of SMARTCOP as a LSM module, the component's processes are allowed to make operations only permitted to the administrator officer – such as packet filtering and application cancellation. The internal access control mechanisms at the kernel are based in the process identifier (PID) that makes the system call, which will be associated to a specific element. Each function registered by a LSM module, determines which component is making the call from the PID of the associated process. It then, applies the access control constraints taking also into account the parameters of the system call. Thus, for example, a given element can access its own configuration files but not configuration files from other elements.

Our protection strategy introduces, however, some administration constraints, since officers are not longer allowed to throw system calls which may suppose a threat to the protected component. To solve these constraints, we also presented in [5] a smart card based authentication mechanism, based on secret-key cryptography, which acts as a reinforcement of the kernel-based access control. The objective of this complementary mechanism is twofold. First, it holds to the administrator the indispensable privileges to carry out management and configuration activities just when she verifies her identity through a two-factor authentication mechanism. Second, it allows us to avoid those attacks focused on getting the rights of the administrative entity, such as dictionary-based attacks and buffer overflows.

Nevertheless, and although the authentication mechanism proposed in [5] solves the administration constraints of our approach, it presents important drawbacks. For instance, there is a need for the entities to share a secret-key, and this is a serious disadvantage for the administration officer, who may be in charge of managing such keys. The process of changing or updating the shared secret-key of all the entities, for instance, over the complete set of components of a network will be very awkward, making it even unfeasible when using our authentication mechanism on huge corporation networks with multiple resources to protect. For this reason, and in order to make easier the administration tasks of our protection approach, we extend in this paper our previous authentication mechanism by using a new authentication protocol based on public key cryptography. Indeed, our new proposal solves the administration constraints of SMARTCOP by using a hierarchical structure with several domains, where the nodes of each domain can independently be administrated by using X.509 certificates [12]. Through this new authentication mechanism, some of the previous drawbacks, such as the sharing of the protocol's information, should be more efficiently performed by means of certificate revocation, for example.

The remainder of this paper is organized as follows. We first define in Section 2 the structure and elements for our new authentication proposal, and present the cryptographic protocol intended to solve the administration constraints introduced by the protection approach described above. We then continue in Section 3 by presenting some configuration issues of our proposal and showing the results of an evaluation of the overhead introduced by our approach on a given setup. We finally summarize in Section 4 some related works, and close the paper in Section 5 with a list of conclusions and future work.

2 Smart Card Based Authentication Mechanism

In order to verify the administrator's identity of SMARTCOP, we propose a two-factor authentication mechanism based on the cryptographic functions of a smart card. This mechanism is intended for authenticating the administrator to the LSM modules and holds with the following requirements: (1) the actions must be authorized by the use of a smart card; (2) the smart card only authorizes one action whenever the PIN would be correct; and (3) the LSM module only authorizes the action whenever the smart card response would be valid, i.e., the cryptographic operation is correct.

Let us start the description of our authentication mechanism by introducing the necessary structure and elements for our proposal. We first define the necessary architecture for our authentication protocol as a hierarchical structure with several organizational units, where the network is divided, in turn, in hierarchical domains, and where each domain of the network has several components that must be protected. We name such a component as SMARTCOP Node (SCN). Each domain has moreover a SMARTCOP Server (SCS), and each potential administrator holds a SMARTCOP Card (SCC). These component are briefly described next.

SMARTCOP Server (SCS) – Each SCS owns a cryptographic key pair *master key* and the corresponding certificate. This certificate has been issued by the upper SCS in the hierarchy and identifies the lower SCS as a valid SCS. This certificate is encoded as an X.509 Attribute Certificate [12], where the issuer is the upper SCS master key and the subject is the lower SCS master key. The SCS of domain B can issue certificates authorizing a concrete SCC as an administration of the domain B (similar to the certificates between SCSs). The SCS must usually be managed by the network administration officer of the given domain – or organizational unit. That is, the person who has more knowledge about the network domain and its potential administrators, and, at the same time, the one that has the greatest interest in performing a good administration. This is a key point of the extended authentication proposal, which enables the distribution of the administrative management between domains or organizational units.

SMARTCOP Node (SCN) – Each SCN is a component which has the SMARTCOP LSM module. The security parameters of the LSM module are properly initialized when it is installed. The main parameter is the *Source-of-Authority* (SoA), which is represented by a *master-key*. More precisely, the *master-key* of the top SCS. When an administrator requests a protected action on a given SCN, by using Protocol 1, the SCN verifies the certificate from the SCC. Then, if it comes from a certificate path rooted at the SoA's *master-key*, the operation is accepted.

SMARTCOP Card (SCC) – The SCC is owned by potential administrators. In order to be able to perform administrative tasks on a given domain, the SCC must be authorized (i.e., certified) by the SCS of the domain or an upper one in the hierarchy. Each SCC has a key pair, which has to be certified by a *master-key* (i.e., a key from a SCS). Let us recall that the cryptographic engine of such a smart card is capable of performing several cryptographic functions, such as asymmetric key generation, asymmetric cryptographic algorithms execution, and so on.

The SCC has an *operation PIN* and an *administration password*. The operation PIN is at least six digits long and is used to authorize the protected actions. On the other

hand, the administration password is used to change the operation PIN and other management tasks. The system administrator has three consecutive chances to enter the operation PIN. In the third entry, if the smart card receives an incorrect operation PIN, it blocks itself. The smart card can only be unblocked with the administration password. Again, there are three chances to enter the correct administration password. Otherwise, after the failing of three consecutive wrong administration passwords, the smart card blocks itself and becomes useless.

2.1 Protocol Description

We give here a detailed description of the cryptographic protocol that leads our smart card based authentication mechanism. It starts in Step 1 when the system administrator requests an action to the LSM module. We assume here that action X must be authorized by using the smart card. The LSM module blocks immediately in Step 2a the communication channel between the smart card reader and the LSM module. In this way, we can assure that the data sent between the module and the smart card can neither be sniffed nor tampered. The module also forces to remove the smart card when is not necessary. In Step 2c, the LSM module waits for the smart card insertion, and in Step 4e the LSM module does not proceed until the smart card has been removed. In Step 3 the operation PIN travels in a secure way from the keyboard because the LSM module has blocked the channel between the keyboard and the module itself. Then, LSM sends a NONCE obtained at random and the PIN in step 4c. The smart card returns the digital signature of the NONCE computed with the smart card's private key. The protocol concludes in Step 4g where the LSM module verifies whether the digital signature has been computed properly and the digital certificate is valid.

Protocol 1

1. *The system administrator opens a new console and she requests an action X ;*
2. *LSM receives the request from the console and it does the following steps:*
 - (a) *Block the channel and open a connection with the smart card reader;*
 - (b) *Print a message asking to insert the smart card into the reader;*
 - (c) *While the smart card has not been inserted do;*
 - i. *Detect the insertion of the smart card;*
 - (d) *Print a message asking for the operation PIN;*
3. *The system administrator types the operation PIN in the keyboard;*
4. *The LSM does the following steps:*
 - (a) *Obtain the operation PIN;*
 - (b) *Obtain a NONCE value at random;*
 - (c) *Execute the Procedure 1 inside the smart card by using the operation PIN and the NONCE, and obtain a response μ ;*
 - (d) *Print a message to remove the smart card from the smart card reader;*
 - (e) *While the smart card has not been removed do;*
 - i. *Detect the removing of the smart card;*
 - (f) *if μ is ERROR the LSM does not authorize the action X ;*
 - (g) *else do:*
 - i. *Check if the digital signature has been computed with a public key, which belongs to a certification path rooted at the master key (SoA).*

- ii. Verify the smart card certificate against a valid CRL.
- iii. Verify the digital signature μ with the public key P_K obtained from the smart card certificate, $P_K(\mu) \stackrel{?}{=} H(NONCE)$;
- iv. if the verification is correct the LSM authorizes the action X
- v. if the verification is not correct the LSM does not authorize the action X ;

We show next the procedure that is executed within the smart card (cf. Procedure 1). Through such a procedure, the smart card can validate the operation PIN. Whenever the operation PIN is valid, it computes the digital signature of NONCE with the smart card private key.

Procedure 1 [*PIN*, *NONCE*]

1. Validate the operation PIN;
2. If the operation PIN is correct do:
 - (a) Compute the digital signature of NONCE with the private key S_K ,
 $\mu = S_K(NONCE)$;
 - (b) return μ ;
3. If the operation PIN is no correct return ERROR;

To ensure the proper execution of both Protocol 1 and Procedure 1, we have also considered the protection of the entities and the channels involved in such a process, avoiding attacks such as impersonation and channel data manipulation. First, the LSM module guarantees that the binary file of the console can not be overwritten by anyone (even the security officer), remaining the permissions as read-only. Second, the console's executable is compiled in a static fashion. This allows us to reduce the complexity of the protection's console process, since we do not need to consider extra tasks introduced by the loading of shared libraries and its associated files. It also allows us to centralize and reduce the failure points that could be used by a remote attacker which tries to tamper the console's process. Third, the LSM module also controls that each system call launched by any other process in the system does not interfere the normal execution flow of the console's process, such as keyboard key capture, cancellation, or debugging process system calls.

It is also important to recall that the communication channel can not be manipulated by any opponent, since the LSM mediates between the system calls related with the communication channels and the entities that take part within the protocol. Furthermore, and as pointed out in [5], the LSM module does not need to be directly protected since we can assume the kernel environment as a trusted area – since it is mandatory for the kernel security model of any modern operating system.

3 Configuration and Performance Evaluation

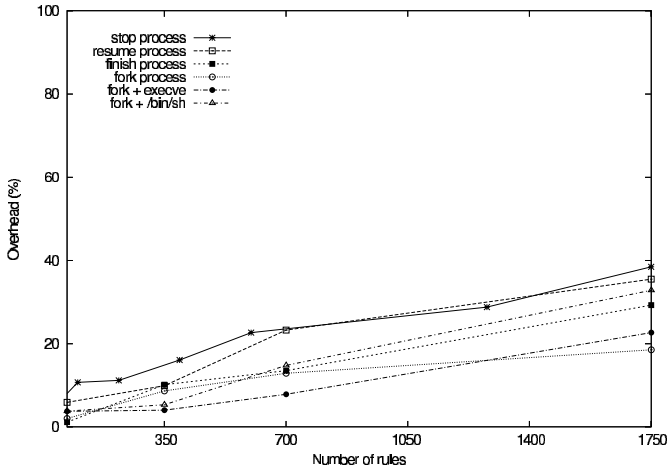
In order to define the objects and resources to protect, SMARTCOP can actually be configured through a set of security rules. Each rule defines an *action* in $\{deny, accept\}$ that applies over a set of *condition* attributes, such as *user_id* (*UID*), *process_id* (*PID*),

device, i-node, etc. We can also define, through these security rules, either open or closed default policies. The complete set of rules are stored in a set of configuration files that are loaded at boot time through the *proc file system*. The *proc file system* (*procfs*) is a special virtual file system in the Linux kernel which allows user space programs to access kernel data structures.

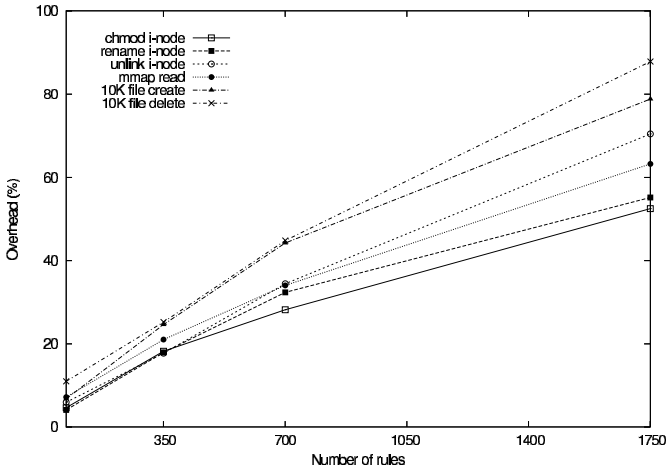
Up to now, we have defined different points through *procfs* for configuring the protection of the three basic levels of protection stated in Section 1. More specifically, we have defined the following entries: *iperms*, *iren*, *isetattr*, *iunlink*, *tcreate*, and *kill*. The four first labels refer to i-node related operations (resp., i-node permission verification, i-node renaming, i-node permission changing, and i-node removing). They can be used not only for the protection of file resources, but also for the protection of communication operations through, for example, sockets and pipes. The last two labels (i.e., *tcreate* and *kill*) are related to the managing of processes (such as creation, suspending, resuming, termination, and spawning of processes). Through these configuration points, we conducted several tests steered towards measuring the penalty introduced by the installation of SMARTCOP as a LSM module, over the normal operation of the system. The tests and benchmarks were based on LMBench [8] and other related administration tools. The evaluation was carried out on a single machine with an Intel-Pentium M 1.4 GHz, with 512 MB of RAM memory and an IDE hard disc of 5400 rpm, running a Debian GNU/Linux operating system and ext3 file system.

During these tests, we measured the overhead of our approach with an instance of SMARTCOP configured with a closed default policy and loaded with different protection rules. More specifically, each configuration point was charged with a set of auto-generated *accept* rules, initially empty, and which incremented to more than three hundred rules. Therefore, a progressive set of auto-generated *accept* rules from zero to more than one thousand rules was globally loaded. We consider that the overlaps between rules, related with the single operations we measured during these tests, represent the worst case scenario we can actually measure. We show in Figure 1(a) and Figure 1(b) the overhead evolution of some actions that we consider representative regarding the set of configuration points we described above.

The first three curves we show in Figure 1(a) represent the overhead evolution of the system call *kill* when we, resp., suspended, resumed, and cancelled a set of processes under the different load of rules. Notice that such actions, especially when suspending processes, reported an acceptable penalty (aprox. a 40% for a global average of almost two thousand rules). The other three curves in Figure 1(a) represent the overhead evolution of the set of operations related to the creation of processes through *fork()*, *fork()+exec()* and *fork()+bin/sh*. Notice that the two first operations supposed a penalty even lower (aprox. a 20% for the highest average of rules); and the third operation remained close the 30% for the same number of rules. Similarly, the results we show in Figure 1(b) are related to the evolution of operations for the managing of i-nodes (such as files, pipes, and sockets' managing). We can appreciate in these results, however, that the penalty introduced by SMARTCOP for the managing of i-nodes seems much higher than the overhead introduced for the managing of processes – it even reached more than an 80% in the operations of file creation and removing. However, we consider that these differences are reasonable, taking into account that there was an overlap between



(a) Processes tests



(b) Filesystem and communication tests

Fig. 1. Performance evaluation of SMARTCOP

processes protection’s rules and i-nodes protection’s rules – expressly introduced during our experiments to simulate the worst case scenario. This overlap between rules definitively exercises a bad influence on the measured i-node operations, compared to the processes operations, and it explains the differences between both results.

4 Related Work

There are two main approaches to safely execute processes with special privileges on modern operating systems. A first approach is the creation of restricted environments, in which the processes will be executed and controlled outside the trusted system space.

In [6], for instance, we can find a traditional mechanism for the creation of restricted environments within Unix setups. These proposals require, however, a replicated file system tree for the protected environments. Hence, the administrator in charge of the system must reproduce the original file system tree to include, for example, shared libraries or configuration files, and copy them to the new path. Other disadvantage of these proposals is that they do not guarantee the correct execution flow of processes, i.e., the behavior of a given process can be modified by using, for example, a buffer overflow. Hence, the attacker can overwrite the configuration or log files of such a process by simply using an arbitrary code execution attack – since these files remain in the same environment of the protected security component process.

A second approach, as the one presented in this paper, is to apply a kernel based access control to outgoing system calls. In [7] and [9], for instance, two similar proposals to ours are presented. The main goal behind these two proposals is to reinforce the complete system by controlling the system calls and ensuring which process or user does the system call and against what it will be done. The ability to control the access to the resources allows to protect system's elements and to avoid that nobody (including an attacker with administrator privileges) can disable them. Nevertheless, both approaches differ from ours in a number of ways. First, and to our best knowledge, neither [7] nor [9] do not address the management of administration constraints, as our proposal does through the two-factor authentication mechanism we present in Section 2. Second, our approach, entirely based on the *Linux Security Modules* (LSM) framework [11], guarantees the compatibility with previous applications and kernel modules without the necessity of modifications. Both [7] and [9] require the rewriting of some features of the original operating system's kernel to properly work. This situation may force to recompile existing code and/or modules in order to obtain the new security features. Although it exists a LSM-based prototype for the approach presented in [7], it does not seem to be actively maintained for the current Linux-2.6 kernel series.

5 Conclusions

We have presented in this paper an access control mechanism specially suited for the protection of network security components, such as *firewalls* and *Intrusion Detection Systems*. Whenever one of these components, or one of its elements, is compromised by a remote attacker, it may lead her to obtain the full control of the network [2]. The protection of these components is not easy, specially when dealing with distributed setups, made up of different elements distributed over a complex network. Like for example, the attack prevention platform presented in [3]. The solution we provided proposes the protection of the components by making use of the *Linux Security Modules* (LSM) framework for the Linux kernel over GNU/Linux systems [11]. The developed mechanism works by providing and enforcing access control rules at system calls, and is based on a protection module integrated into the operating system's kernel, providing a high degree of modularity and independence between elements. Furthermore, the use of a complementary authentication method, based on smart card technology and a public-key cryptographic protocol, allows us to properly verify administrator's actions when officers need to do administration tasks. This additional enhancement also allows us

to prevent some logical attacks against the protection mechanism itself (e.g., password forgery). The integration of our approach on a normal system setup proved, moreover, a good degree of transparency to the administrator in charge, and a reasonable performance penalty for the managing of processes, files, and communication resources.

As a future extension of our work, we are considering improving the customizing of policies. Up to now, the specific policy that is enforced by our protection module is loaded at boot time through the *proc file system* (procf). We are planning to extend this feature to add the possibility of using text-based configuration files and the reload of policies at runtime. We are also considering to continue our study to address the security of the system from an intrusion tolerance point of view [1].

References

1. Y. Deswarte, L. Blain, and J. C. Fabre. Intrusion tolerance in distributed computing systems. *IEEE Symposium on Security and Privacy*, pages 110–121, Oakland, CA, USA, 1991.
2. D. Geer. Just How Secure Are Security Products? *IEEE Computer*, 37(6):14–16, 2004.
3. J. García-Alfaro, F. Autrel, J. Borrell, S. Castillo, F. Cuppens, and G. Navarro. Decentralized publish/subscribe system to prevent coordinated attacks via alert correlation. *6th Int. Conf. on Information and Communications Security*, 223–235, Spain, 2004.
4. J. García-Alfaro, S. Castillo, G. Navarro, and J. Borrell. ACAPS: An Access Control Mechanism to Protect the Components of an Attack Prevention System. *Journal of Computer Science and Network Security*, 5(11):87-94, 2005.
5. J. García-Alfaro, S. Castillo, J. Castellà-Roca, G. Navarro, and J. Borrell. Protection of Components based on a Smart-card Enhanced Security Module. *1st International Workshop on Critical Information Infrastructures Security, Information Security Conference (ISC'06)*, Samos, Greece, 2006.
6. P. Hope. Using Jails in FreeBSD for Fun and Profit. *Login; The Magazine of Usenix & Sage*, 27(3):48–55, 2002.
7. P. Loscocco and S. Smalley. Integrating Flexible Support for Security Policies into the Linux Operating System. *11th FREENIX Track: 2001 USENIX Annual Technical Conference*, USA, 2001.
8. L. McVoy. LMBench, Portable Tools for Performance Analysis. *1996 USENIX Annual Technical Conference*, USA, 1996.
9. A. Ott. The Role Compatibility Security Model. *7th Nordic Workshop on Secure IT Systems (Nordsec 2002)*, Karlstad University, Sweden, 2002.
10. J. Viega, and G. McGraw. *Building Secure Software - How to Avoid Security Problems the Right Way*. Addison-Wesley, 2002.
11. C. Wright, C. Cowan, S. Smalley, J. Morris, and G. Kroah-Hartman. Linux Security Modules: General Security Support for the Linux Kernel. *11th USENIX Security Symposium*, USA, 2002.
12. ITU-T. The Directory: Public-key and attribute certificate frameworks. ITU-T Recommendation X.509, 2000.

On the Existence of Related-Key Oracles in Cryptosystems Based on Block Ciphers

Ermaliza Razali and Raphael C.-W. Phan

Information Security Research (iSECURES) Lab,
Swinburne University of Technology (Sarawak Campus), Malaysia
{erazali, rphan}@swinburne.edu.my

Abstract. The notion of a related-key attack (RKA) was formally introduced by Biham in 1993. It is essentially more of an attack model rather than a specific type of attack in that it considers what sort of oracles are available to the attacker. In this case, the attacker has access to related-key (RK) oracles, i.e. he is able to have encryptions performed on plaintexts of his choice, keyed by two or more unknown but related keys. The feasibility of this attack model is at times debated mainly because the assumption that an attacker would have access to RK oracles may be too strong to really exist in practice. Hence, attacks on block ciphers in this RKA model have commonly not been regarded on the same level of significance of those not requiring RK oracles. A good example is the AES. It is generally accepted that the best known attack is a non-RKA by Gilbert and Minier in 2000, although it applies to less rounds compared to the best known RKA on AES by Biham et al. that applies to more rounds. It is our aim in this paper to show how RK oracles exist in various block cipher based cryptosystems. The gist is to think outside the box, i.e. to note that a block cipher is often an underlying primitive within a larger cryptographic construct, thus it is only natural to evaluate the block cipher security in this setting and not as a standalone primitive. In doing so, we formally introduce the notion of related-key multiplicative differentials, and related-key compositionally differentials. We also consider the existence of RK oracles in PGV-type hash functions, message authentication codes, recent authenticated encryption modes and cases of key-exchange protocols not previously mentioned in literature.

1 Introduction

The notion of related-key cryptanalysis of block ciphers was first formalized by Biham [2] as a cryptanalytic tool for analyzing block ciphers although earlier work such as [38,20] exploited the existence of related keys to aid cryptanalysis of specific ciphers. This notion was later put into a more theoretical setting by Bellare and Kohno [1], and further by Lucks [24]. A *related-key attack* (RKA) assumes that the attacker is capable of obtaining the encryption of some plaintexts under two or more different but unknown keys which have some relationship between them. In essence, a RKA is more of an attack model (other attack models for block ciphers include known-plaintext and chosen-plaintext models etc)

rather than a particular type of attack (e.g. differential attack, linear attack). This is so because similar to other attack models, a RKA considers what sort of oracle is available to the attacker. In the particular case of this, the attacker has access to a *related-key (RK) oracle* [1] which will be defined more formally in Section 2.

It is common folklore in the cryptologic research community that RKAs seem unrealistic in the sense that it is infeasible to find scenarios where block ciphers allow for RK oracles, i.e. when ciphers are used that perform encryptions under two or more secret keys related in some way known to an adversary. This is the reason why interest in RKAs is mostly confined to hardcore block cipher designers and cryptanalysts, and why related-key attacks on block ciphers are sometimes treated with skepticism, e.g. the current best attacks (in terms of being applicable to the most number of rounds) on the AES [25] are related-key ones [9,4], but it is generally accepted that the best feasible one is a non-RKA [10]. In fact, this is also true for several other ciphers [16,22,4,3] where RKAs are the best known results. Thus if we could remove the conventional mindset that RKAs are unrealistic, then in addition to giving more significance to these attacks, it would also allow to have a fairer comparison of the security of ciphers against attacks that would then include related-key ones as equally significant, and encourage designers to give more emphasis to key schedule design. From a different perspective, if we consider the effects of attacks irrespective of what attack model they are in, then while non-RKAs could demonstrate weaknesses in the design of any cipher component, existence of RKAs almost always demonstrate weaknesses in the design of the key schedule thereby allowing to exactly pinpoint this part of the cipher for more strengthened designing. Further, [1] proved that there exist some classes of RKAs that no cipher can resist, i.e. if such attacks are disregarded by a cipher designer or implementer as infeasible, but then somehow an attacker manages to access such RK oracles, then the cipher will completely be broken.

In this paper, we emphasize that we are not proposing specific RKAs on any cipher. Doing so would be similar to past work (see [22,4] for some examples), and again the question would therefore arise as to why it is feasible in practice to have RK oracles. Instead, we will attempt to address this latter folklore by showing that various scenarios exist in block cipher based cryptosystems that allow for RKAs to be mounted, i.e. RK oracles exist. The basic approach is to *think outside the cipher box*, i.e. to note that block ciphers are often used as underlying primitives in larger cryptographic constructs, e.g. in modes of operation [26], message authentication codes (MACs) [13], recent authenticated encryption modes [27,29], stream ciphers, hash functions [35], authentication and key-exchange protocols, and commercial security systems like those in software (e.g. PGP, WinZip), hardware like tamper-resistant devices (e.g. IBM 4758 cryptoprocessor) or smartcard-based applications. In such settings, the block cipher is either interacting with other primitives within the larger construct, or with one or more copies of itself (e.g. triple-DES is a composition of 3 copies of DES keyed by multiple secret keys). Thus, to be complete it is more reasonable

to gauge the security of a block cipher in this setting, rather than merely considering it as an independent standalone primitive. And in essence, this latter is precisely what the RKA model does because it simultaneously considers two or more copies of a cipher keyed by different but related keys. So the first step towards dispelling the folklore is to *not* view block cipher security as just that of a standalone primitive.

In fact, for some primitives other than block ciphers, researchers are also realizing the importance of moving away from analyzing the security of a primitive existing on its own (standalone security) and towards analyzing the primitive as part of a bigger construction. This is because it has been demonstrated that primitives proven secure when in isolation no longer remain so when composed with different proven primitives [11,17]. Therefore, results such as those in the universally composable (UC) framework by Canetti [7] and reactive simulatability by Pfitzmann and Waidner [30] demonstrate the significance of the approach of composable security. Namely, when a primitive is proven secure under such frameworks, then by a composition theorem we are guaranteed that the primitive will still be secure when plugged as an underlying building block into any arbitrary larger cryptographic construction and further we retain the guarantee that the latter construct is also secure.

RELATED WORK. Considering scenarios for RKAs is not new. In fact, this was necessary in justifying why it is worth to analyze the security of block ciphers against RKAs in the first place. Under the context of mechanical rotor machines, if the operator sets rotors incorrectly, then there is possibility of obtaining encryptions under related keys [8]. Meanwhile, the basic idea of exploiting related keys to block cipher-based hash functions appears to be first mentioned in [34], though it was not called as such nor put in the context of block ciphers. It was only in [2] that the RKA model was formally introduced in the context of block cipher security, and it was also here that was mentioned how RKAs on block ciphers are applicable when using block ciphers in hash functions. [15,16] later described how key-exchange protocols that do not provide integrity, or that transmit a key schedule's salt in the clear, or updated keys using a known function, would allow for RKA scenarios. They also mentioned that related keys could be obtained from several key-exchange sessions or by attacking n -party key-exchange protocols. [32] described how a tamper-resistant cryptoprocessor allowed scenarios for related-key XOR differential and related-key slide attacks. Elsewhere, though not in context of block ciphers as underlying primitives but where hash functions are instead, [19] discussed how block ciphers based on hash functions could be attacked faster than monolithic block ciphers designed from scratch, while [18] described attacks on a message authentication code HMAC if it had some hash functions as primitives.

OUR RESULTS. We first classify types of RK oracles using the notion of related-key-deriving (RKD) function in [1]. In doing so, we introduce the notions of RK multiplicative differential and RK compositional differential. We also formalize the notion of the RK slide. Throughout this paper, we consider block cipher-based cryptosystems and show scenarios where the way in which the underlying

block cipher is used allows for RK oracles. In particular, we show that though confidentiality-only modes and authenticated encryption modes do not allow for RK oracles, one recently proposed authenticated-only mode does. There are also further subtle points with key-exchange and key management protocols that may allow RK oracles. In our concluding section, we give some interesting open problems related to the interaction of multiple types of RK oracles and the relationships between their RKD functions.

2 Preliminaries

A block cipher E is defined as a map $E : \mathcal{K} \times \mathcal{D} \rightarrow \mathcal{D}$, where $\mathcal{K} = \{0, 1\}^k$ is the secret key space, $\mathcal{D} = \{0, 1\}^n$ is the text (plaintext/ciphertext) space, k is the secret key length in bits, n is the blocksize in bits. We use $E_K(D)$ as shorthand for $E(K, D)$.

A *related-key-deriving* (RKD) function $\phi \in \Phi$ is a map $\phi : \mathcal{K} \rightarrow \mathcal{K}$, where Φ is a set of all functions mapping \mathcal{K} to \mathcal{K} . Given E and $K \in \mathcal{K}$, we define the *related-key (RK) oracle* $E_{RK(\cdot, K)}(\cdot)$ as an oracle taking two arguments: a function $\phi : \mathcal{K} \rightarrow \mathcal{K}$ and an element $P \in \mathcal{D}$, and that returns $E_{\phi(K)}(P)$; where $RK(\phi, K) = \phi(K)$. An attack based on exploiting access to the RK oracle $E_{RK(\phi, K)}(\cdot)$ for $\phi \in \Phi$ is called a Φ -restricted¹ RKA. Examples of Φ considered so far in literature [1] include related-key (RK) additive differential Φ_k^+ , i.e. the set of functions $K \rightarrow K + i \bmod 2^k$ for $0 \leq i < 2^k$; and RK exclusive-OR differential Φ_k^\oplus the set of functions $K \rightarrow K \oplus \Delta$ for $\Delta \in \{0, 1\}^k$. These are important classes of RKD functions because they encompass almost all of the RKAs considered in literature. [1] are also the first to consider *multiple-type* RKAs where the attacker has access to more than one type of RK oracle, namely they considered the case where the RKD function is $\Phi_k^+ \cup \Phi_k^\oplus$ and gave an impossibility result, i.e. no cipher is resistant to RKAs given access to the $\phi \in \Phi_k^+ \cup \Phi_k^\oplus$.

We formalize the notion of *RK slide* Φ_k^{\leftrightarrow} [2,31], i.e. the set of functions $K \rightarrow K'$ where $K = K_1 || K_2 || \dots || K_l$ and $K' = K'_1 || K'_2 || \dots || K'_l$ such that $K_i = K'_{i+s}$ for $s \in \{1, \dots, l-1\}$. We also introduce the notion of *RK multiplicative differential* Φ_k^\times , i.e. the set of functions $K \rightarrow K \times \Delta \bmod 2^k$ for $\Delta \in \{0, 1\}^k$, inspired by the notion of multiplicative differentials [5]. Such a notion for related keys has not been considered before. Another notion that we formally introduce is *RK compositional differential* Φ_k^f the set of functions $K \rightarrow f(K)$ where $f = \phi_l(\dots(\phi_2(\phi_1(\cdot))))\dots$ is a composition of l different RKD functions $\phi_1, \phi_2, \dots, \phi_l \in \Phi_k^f$.

3 Hash Functions Based on Block Ciphers

We take a soft first step by considering a typical block-cipher based construction that is often cited for allowing RK oracles: block-cipher based hash functions. A hash function $H : \{0, 1\}^* \rightarrow \{0, 1\}^m$ takes a message of arbitrary

¹ This is called Φ -transforming in [24].

length $M \in \{0, 1\}^*$ as input and produces an output $H(M) \in \{0, 1\}^m$ of a fixed length string of m bits, known as a hashcode, hash-value or simply hash. Constructing a hash function based on a block cipher [35] is often more desirable especially in space-constrained environments since this would only require a single implementation (i.e. the block cipher) for both cryptographic primitives, namely a block cipher and a hash function. There are several methods in which a block cipher can be used as the underlying primitive for a hash function including Davies-Meyer, Miyaguchi-Preneel and Matyas-Meyer-Oseas. These are commonly known as PGV-type schemes, and were considered by Preneel et al. [35]. In particular, they analyzed 64 schemes whose compression function $f(\cdot, \cdot)$ is of the form:

$$f(K, P) = E(K, P) \oplus FF \tag{1}$$

where $E(\cdot, \cdot)$ is a block cipher, P , K , and FF can be chosen from the set of $\{\text{constant value} = V, \text{message block} = X_i, \text{hashcode} = H_{i-1}, \text{the XOR of plaintext and hashcode } X_i \oplus H_{i-1}\}$.

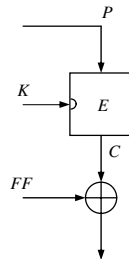


Fig. 1. General structure of the hash compression function

It is commonly known [33] that the Davies-Meyer scheme allows RKAs to be mounted on the block cipher primitive within a block cipher-based hash function. To complete the picture, we will analyze **all** 64 PGV schemes to show which ones besides the Davies-Meyer cause RK oracles. We use the same scheme notations as in [35].

Table 1. 12 of the 64 schemes extracted from Table 1 in [35]

Choice of FF	Choice of K	Choice of P			
		X_i	H_{i-1}	$X_i \oplus H_{i-1}$	V
V	X_i	-	scheme 13 (Rabin's scheme)	scheme 25	-
X_i	X_i	-	scheme 15	scheme 29	-
H_{i-1}	X_i	scheme 5	scheme 19 (Davies-Meyer)	scheme 33	scheme 44
$X_i \oplus H_{i-1}$	X_i	scheme 9	scheme 21	scheme 37	scheme 46

Consider first the class of PGV schemes in Table 1. There are 12 such schemes including Rabin's scheme [36] and the popular Davies-Meyer scheme. These schemes have a common attribute i.e. they feed each block X_i of the message

M as the key K to the underlying block cipher E . Thus if we choose the message blocks being input to the hash function such that they differ in some way, then we obtain executions of the underlying block cipher under these different keys. We then obtain related-key queries on the block cipher, i.e. a RK oracle. In particular, two message blocks, X_i and X'_i that differ with the relationship $\Delta = X_i \oplus X'_i$ would cause the underlying block cipher E to be keyed under keys with difference Δ , i.e. we get $E_{X_i}(\cdot)$ and $E_{X_i \oplus \Delta}(\cdot)$.

The same principle applies to 16 other schemes in Table 2 because if only the message blocks X_i and X'_i differ but all previous messages blocks X_j and X'_j for $j < i$ are equal, then previous hashcodes H_j and H'_j for $j < i$ would also be equal and thus cancel out in the computation of the difference $\Delta = X_i \oplus X'_i$. So we similarly get $E_{X_i}(\cdot)$ and $E_{X_i \oplus \Delta}(\cdot)$.

Table 2. 16 of the 64 schemes extracted from Table 1 in [35]

Choice of FF	Choice of K	Choice of P			
		X_i	H_{i-1}	$X_i \oplus H_{i-1}$	V
V	$X_i \oplus H_{i-1}$	scheme 2	scheme 14 (Bitzer’s scheme)	scheme 27	scheme 41
X_i	$X_i \oplus H_{i-1}$	scheme 4	scheme 17	scheme 31	scheme 43
H_{i-1}	$X_i \oplus H_{i-1}$	scheme 7	scheme 20 (LOKI mode)	scheme 35	scheme 45
$X_i \oplus H_{i-1}$	$X_i \oplus H_{i-1}$	scheme 11	scheme 23	scheme 39	scheme 48

Similarly, schemes in Table 3 feed the previous hashcode H_{i-1} as the key K to the underlying block cipher E with feedforward FF in the form of X_i or $X_i \oplus H_{i-1}$. Hence choosing the feedforwards in a manner that they differ in some way affects the resulting hashcodes, and we can obtain encryptions of the underlying block cipher under these different keys. In particular, two feedforwards, $X_i \oplus H_{i-1}$ and $X'_i \oplus H_{i-1}$ that differ with the relationship $\Delta = X_i \oplus X'_i$ would cause the underlying block cipher E to be keyed under keys with difference Δ , i.e. we get $E_{H_{i-1}}(\cdot)$ and $E_{H_{i-1} \oplus \Delta}(\cdot)$.

Table 3. 8 of the 64 schemes extracted from Table 1 in [35]

Choice of FF	Choice of K	Choice of P			
		X_i	H_{i-1}	$X_i \oplus H_{i-1}$	V
X_i	H_{i-1}	scheme 3	scheme 16	scheme 30	scheme 42
$X_i \oplus H_{i-1}$	H_{i-1}	scheme 10	scheme 22	scheme 38	scheme 47

Using the notion of RKDs, we can generalize this to saying that any two or more message blocks X_i, X'_i, \dots for any $\phi \in \Phi_k^\oplus$ such that $\phi : X_i \rightarrow X'_i$ that are input into a PGV block cipher-based hash function, would cause an RK XOR differential oracle $E_{\phi \in \Phi_k^\oplus}(\cdot)$.

4 Modes of Operation

A block cipher encrypts plaintext in fixed n -bit blocks, where n is the block size. Nonetheless, a message to be encrypted is not always restricted to n bits,

therefore, for messages which exceed n bits, they need to be partitioned into n -bit blocks before being encrypted n bits at a time. There are several modes of which a block cipher can operate, the most standard ones being the electronic codebook (ECB) mode, the cipher block chaining (CBC) mode, the cipher feedback (CFB) mode, the output feedback (OFB) mode, and the counter (CTR) mode [26]. Note that these five modes of operation only provide confidentiality. Recently, more modes that provide the additional authentication feature have been proposed and recommended by NIST [27,29] including CCM (for authenticated encryption) and GCM (for parallelizable authenticated encryption).

We will consider for all these if any of them allow for RK oracles. We remark that previous work e.g. [21] have considered related-key attacks directly on modes of operation, but this is different from our treatment in this section that gives justification why the use of modes of operation allows for RK oracles against the underlying block cipher.

For *confidentiality-only* modes and for recently proposed NIST *authenticated encryption* modes, we answer in the negative. The proposed NIST *authenticated-only* modes give similar results. An earlier proposal, however, allows RKAs.

CONFIDENTIALITY MODES. We give below an example to show how these confidentiality modes do *not* allow for RK oracles. Let us consider the **CBC** mode defined by $C_1 = E_K(P_1 \oplus IV)$ and $C_j = E_K(P_j \oplus C_{j-1})$ for $j = 2, \dots, i$ where P_i , C_i are the plaintext-ciphertext blocks, IV denotes the initialization vector and K is the encryption key of the underlying block cipher. Observe that modifying (e.g. flipping a bit in) either one of these parameters (P_i, IV, C_i) will *not* affect K . In addition, the key, K which is being fed to the encryption function, E is fixed for the entire operation. Thus, a RK oracle does *not* exist due to the aforementioned properties.

The same principle applies to the other confidentiality modes, i.e. ECB, CFB, OFB and CTR [26]. Hence, standard confidentiality modes do *not* allow RK oracles.

AUTHENTICATED ENCRYPTION MODES. [27] specifies a block cipher mode of operation called Counter with Cipher Block Chaining-Message Authentication Code (**CCM**) which provides both confidentiality and authenticity of data. However, the implementation of CCM limits the total amount of data be encrypted with only a single key, K . Furthermore, K is used for both the CTR and CBC-MAC modes within CCM, and as mentioned earlier, the CTR and CBC modes do not allow for RK oracles. Similarly for [29], the proposed draft for Galois/Counter Mode (**GCM**) employs a variant of the CTR mode and a hash function to cater for confidentiality and authenticity of data, respectively. Note also that there is only one key being used for this mode, which is the key to the underlying block cipher. In conclusion, both the CCM and GCM modes do *not* allow for RK oracles.

AUTHENTICATED MODES. [28] detailed a cipher-based MAC (**CMAC**). A MAC can be viewed as a mode of operation that provides authentication only. CMAC is essentially the One-key CBC-MAC (OMAC) [12]. CMAC uses two subkeys

namely K_1 and K_2 which are generated from the key, K and are fixed for any invocation of CMAC under K . The generation process of the subkeys is as follows:

1. Let $L = E_K(0^b)$.
2. If $MSB_1(L) = 0$, then $K_1 = L \ll 1$;
Else $K_1 = (L \ll 1) \oplus R_b$.
3. If $MSB_1(K_1) = 0$, then $K_2 = K_1 \ll 1$;
Else $K_2 = (K_1 \ll 1) \oplus R_b$.
4. Return K_1, K_2 .

Here, $MSB_s(X)$ denotes a bit string consisting of the s left-most bits of the bit string X while R_b , which is made public, denotes a bit string where b is the block size and for $b = 128$, $R_{128} = 0^{120}10000111$ and for $b = 64$, $R_{64} = 0^{59}11011$. K_1 is used to mask the final message block if M is a positive multiple of the block size but if the final block is not a positive multiple of the block size, it is padded with a single '1' bit followed by the minimum number of '0' bits and K_2 is used instead for the masking of this padded block. Note also that from the subkeys generation process, $K_2 = (K_1 \ll 1) \circ R_b$, where \circ denotes \oplus operation that is conditional on $MSB_1(K_1)$. We can further represent as $K_2 = f(K_1)$ for $f = \phi_2(\phi_1(\cdot))$ and $\phi_2 \in \Phi_k^\circ(\cdot)$, $\phi_1 \in \Phi_k^{\ll}(\cdot)$. This appears to provide an RK compositionally differential oracle $E_{\phi \in \Phi_k^f}(\cdot)$. However, since the keys K_1 and K_2 do not actually key the underlying block cipher but instead are used to mask the final plaintext message block, this cannot be exploited to mount RKAs on the block cipher. As a side note, this does provide a non-related-key compositionally differential oracle but that is beyond our scope here.

Another recently proposed mode is the Two-Key CBC MAC (**TMAC**) in [23] by the same designers of [12]. Similar observations can be made on this, i.e. though it appears to exhibit an RK multiplicative oracle $E_{\phi \in \Phi_k^\times}(\cdot)$ since multiplicatively different keys may exist, however this does not result in an RK oracle since the keys do not key the block cipher but are used to mask the final message block.

Finally, we consider Randomized Message Authentication Code (**RMAC**) which was initially proposed as an NIST draft [13] in the effort to standardize authentication modes. RMAC is basically the CBC mode [26] keyed by a secret key K_1 , followed by a randomizing component wrapped around a block cipher keyed by another key K_2 and involving a public randomizer R that changes with each RMAC computation. In more detail, pad the message M to be a multiple of n bits where n is the blocksize of the underlying block cipher, split the padded M into a sequence of l n -bit message blocks M_i for $i = 1, \dots, l$, i.e.

$$\begin{aligned}
 H_1 &= E_{K_1}(M_1), \\
 H_i &= E_{K_1}(M_i \oplus H_{i-1}); (2 \leq i \leq l), \\
 RMAC(M) &= E_{K_2 \oplus R}(H_l).
 \end{aligned}$$

As an example, Figure 2 illustrates the case for RMACing a 2-block message $M = M_1 || M_2$. Just by looking at the above equations, since the randomizer

R varies with each RMAC computation, then we get a different key $K_2 \oplus R$ keying the encryption of the last block every time. This is clearly an RK XOR differential oracle $E_{\phi \in \Phi_k^\oplus}(\cdot)$, for any distinct pair of R and R' corresponding to any pair of RMAC computations.

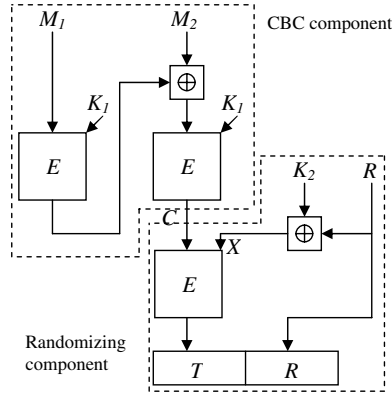


Fig. 2. The RMAC mode

In fact, we can even choose the RK XOR difference Δ . In more detail, if R corresponds to one RMAC computation, we can tweak the randomizer R' of the other RMAC computation into $R'' = R' \oplus (R \oplus R' \oplus \Delta) = R \oplus \Delta$. Thus, the key difference would be $(K_2 \oplus R) \oplus (K_2 \oplus R'') = (K_2 \oplus R) \oplus (K_2 \oplus R \oplus \Delta) = \Delta$.

5 Key Exchange Protocols and Key Management

Key-exchange protocols [6] allow two or more parties to establish a shared secret cipher key for subsequent use with a block cipher; and some of these protocols typically use block ciphers (among others) as primitives. We first discuss previously known cases and then give somewhat new idea sketches on how several types of these protocols allow RK oracles.

WITHOUT INTEGRITY. It has been known [15] for some time that key-exchange protocols that do not provide for key-integrity would allow for RKA scenarios. For instance, if the key-exchange protocol allows any attacker to flip bits in the key even without the knowledge of the key itself, then clearly a Φ_k^\oplus -restricted RKA scenario exists.

[15] has described a related-key scenario for the two-party key distribution protocol (2PKDP) in [37]. This observation also applies to the three-party variant (3PKDP) in [37]:

- $A \rightarrow S : A, B, N_{as}$
- $S \rightarrow A : MAC_{K_{as}}(N_{as}, N_{sa}, B) \oplus K_{ab}, N_{sa}$.
- $B \rightarrow S : B, A, N_{bs}$
- $S \rightarrow B : MAC_{K_{bs}}(N_{bs}, N_{sb}, A) \oplus K_{ab}, N_{sb}$.

where $N_{as}, N_{bs}, N_{sa}, N_{sb}$ are nonces, K_{ab} is a short-term session secret key for a block cipher and K_{as}, K_{bs} are long-term shared secret keys between A and S , and B and s , respectively. From the messages sent by S to A and to B , we notice that an attacker has an RK XOR differential oracle $E_{\phi \in \mathcal{F}_k^\oplus}(\cdot)$ similar to the case in 2PKDP as he needs only flip any bits of the message to change K_{ab} .

Another nice example is in the case of key distribution schemes for Pay-TVs. Service providers incorporate Conditional Access Systems (CAS) inside the decoder box and the smart card to prevent unauthorized access paid content. [14] presented a CAS for Pay-TV systems which contains a scrambling algorithm for the paid content, and a key distribution scheme which ensures that descrambling parameters are sent to the correct decoder box and smart card pair. See Figure 3, where R denotes a random number, S denotes the seed of a pseudo-random number generator $G(\cdot)$, X is the broadcast message, IV is the initialization vector and K_s is the scrambling key. It suffices to note that upon receiving X , the smart card finally computes Y_i which is later transmitted in the clear to the decoder box. Using X and Y_i , the decoder box finally computes S which is the seed to a pseudo-random generator used to key the descrambling algorithm. Hence, an adversary can flip bits of Y_i to Y'_i such that $\Delta = Y_i \oplus Y'_i$. This, in turn, flips the corresponding bits of S , i.e. S now becomes $S' = S \oplus \Delta$. This gives an RK XOR differential oracle $E_{\phi \in \mathcal{F}_k^\oplus}(\cdot)$.

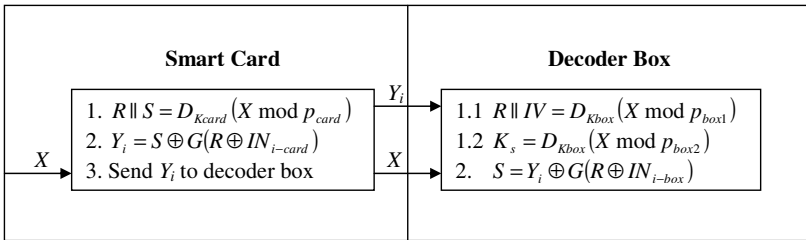


Fig. 3. Operations between the decoder box and the smart card

KEY UPDATE. It is common that shared secret keys need to be updated over time, e.g. in cases of the key being compromised or leaked, or when members of a group (sharing the key) leaves the group or a new member joins. In this case, it is well known [15] that if the shared secret keys are updated using very poorly designed key-update mechanisms, then an attacker would know what the different keys are and therefore we have related-key scenarios. For instance, [16] cites a proprietary implementation that simply updates the keys as $K, K + 1, K + 2$, etc. In fact, even more complex key update mechanisms would allow RK oracles as long as they are deterministic in nature, thereby allowing the attacker to know the relationship between the updated keys [16].

COLLUSION OF KEYS. A block cipher implementation would be used by several users with independent keys K, K', \dots . However, if generation and management

of these keys are done by a centralized trusted third party, then it needs to ensure no obvious dependency exists among these keys, otherwise the collusion of two or more keys would induce RK oracles, i.e. the attacker just needs to gather black box encryptions under these related keys.

KEY CONFIRMATION. In order to ensure all parties really did establish the correct shared key, K ; key-exchange protocols often use an extra key-confirmation step which is basically a 2-move challenge-response-like exchange of encrypted messages between the parties, e.g. A sends $E_K(N)$ while B returns $E_K(N+c)$ for any constant c . This may cause RK oracles, as described in the next paragraph. To further avoid RK oracles, one should not instead vary the key, i.e. if A sends $E_K(N)$ while B returns $E_{K+c}(N)$ then this gives an RK additive differential oracle $E_{\phi \in \Phi_k^+}(\cdot)$.

DIFFIE-HELLMAN-LIKE. Note that for Diffie-Hellman-like key-exchange protocols, the shared secret key is established with equal parts contributed by each party, i.e. $K = g^{xy}$ where g^x is contributed by one party and g^y by another. In such cases, while the parts are being communicated it is possible to modify them by multiplying with a factor α , in which case the key K established by one party would differ multiplicatively from that K' computed by the other. Though a further key confirmation stage would detect this difference, however the very messages used for key confirmation could be exploited e.g. A sends $E_K(N)$ while B returns $E_{K'}(N + c)$ then this gives rise to the RK multiplicative differential oracle $E_{\phi \in \Phi_k^\times}(\cdot)$ since $K' = zK$ for some $z = f(\alpha)$. More of this is explained in the next subsection.

5.1 Related-Key Multiplicative Differential Oracles

In [5], differential cryptanalysis using multiplicative differentials was applied to some ciphers including `xmx`. A multiplicative differential is one that considers differential pairs of the form (x, x') where $x' = \alpha x$, and α is the multiplicative difference between the pair. One then studies the propagation of this difference through the cipher components.

This notion can be extended to the related-key case, i.e. besides having plaintext pairs with multiplicative differences, introduce a pair of related keys (K, K') where $K' = \alpha K$. We then obtain the RK multiplicative differential oracle $E_{\phi \in \Phi_k^\times}(\cdot)$.

The MultiSwap cipher, which is used in Microsoft’s Digital Rights Management system was analysed in [5] for its susceptibility against multiplicative differential cryptanalysis. Their attack uses a multiplicative differential pair (w, w') where $w' = 2w$, which is inserted in the second half of the encryption process. Recall that it is generally assumed that MultiSwap uses independent round keys, k_i for $i = 1, \dots, 10$ i.e. the secret key $K = k_1 || k_2 || \dots || k_{10}$. Thus, observe that the same result is obtained if instead we use a multiplicative differential round key pair (k_6, k'_6) where $k'_6 = 2k_6$. So, instead of having $k_6 \cdot w' = k_6 \cdot (2w) = 2(k_6 \cdot w)$, we have $k'_6 \cdot w = 2k_6 \cdot w = 2(k_6 \cdot w)$. However, the difficult task is to control the multiplicative difference such that it is confined to just one subkey (in this case,

k_6) instead of the entire secret key K , i.e. such that $K = k_1||k_2||\dots||k_6||\dots||k_{10}$ and $K' = k_1||k_2||\dots||k'_6||\dots||k_{10}$. But this is still possible. Consider a key $K = 12_x$ where each hex digit represents a subkey block. Our aim is to obtain another key $K' = \alpha K \bmod 2^7 = \alpha \times 12_x \bmod 128$ in such a manner that only one hex digit is affected by the multiplication. Choosing $\alpha = 12_x$, we have $K' = \alpha K \bmod 2^7 = 12_x \times 12_x \bmod 128 = 90_x \bmod 128 = 10_x$. Hence, only one digit has been changed due to the multiplication. Another example is to regard every n bits as a subkey block. Say we have a key $K = 00000100_2 = 4_{10}$ and a related key $K' = \alpha K \bmod 2^8 = 109_{10} \times 4_{10} \bmod 256 = 436_{10} \bmod 256 = 180_{10} = 10110100_2$, where $n = 4$ and $\alpha = 109$. Thus, the two subkeys are $K = (K_1, K_2) = (0000_2, 0100_2)$ and $K' = (K'_1, K'_2) = (1011_2, 0100_2)$, respectively; and they differ multiplicatively in only one subkey block.

6 Conclusion and Open Problems

We have put forth the case that RK oracles exist in several block cipher based cryptosystems, and thus RKAs on the underlying block ciphers are not as infeasible as conventionally thought. Further, the RKA model rightfully considers the security of a block cipher not just as a standalone primitive but it does so in the context of the cipher within a larger cryptographic construct. Our notions of RK multiplicative differential and RK compositional differential may be of independent interest. It is an interesting open problem to consider RKAs on specific block ciphers that exploit such oracles, since we have demonstrated in this paper that these oracles do exist.

Another point is not just to consider security against a single type of attack in isolation, but to consider access to RK oracles of different types, all of which are available simultaneously. Bellare and Kohno [1] took the first step in this direction by considering the case of $\Phi_k^+ \cup \Phi_k^\oplus$ -restricted RK oracles. In fact, since we are already assuming the RKA model where RK oracles exist, then there is not much more assumption to be made whether we consider RK XOR and RK additive differential oracles to exist independently, or simultaneously. Of course, we should also exercise caution when interpreting any provable security result with respect to RK oracles. For instance, a cipher proven secure against the Φ_k^+ -restricted RK oracle does not imply that if we just restrict to this type of RK oracle, the cipher cannot be broken by RKAs. This is intuitively due to the fact that relationships exist between different RKDs, e.g. $\phi_k^+ = \phi_k^\oplus$ for cases where no carries are caused by ϕ_k^+ or when the carry is due to the most significant bits; or $\phi_k^{<} = \phi_k^x$ for the case where the former shifts left by 1 while the latter has the multiplicative factor $\alpha = 2$.

References

1. Bellare, M., and Kohno, T. A Theoretical Treatment of Related-Key Attacks: RKA-PRPs, RKA-PRFs, and Applications. *Advances in Cryptology - Eurocrypt'03*, LNCS 2656, Springer, pp. 491–506 (2003). Full version available at <http://www-cse.ucsd.edu/users/tkohno/papers/RKA/rka.pdf>.

2. Biham, E. New Types of Cryptanalytic Attacks Using Related Keys. *J. Cryptology* 7, 4, Springer, pp. 229–246 (1994).
3. Biham, E., Dunkelman, O., and Keller, N. A Related-Key Rectangle Attack on the Full KASUMI. *Advances in Cryptology - Asiacrypt'05*, LNCS 3788, Springer, pp. 443–461 (2005).
4. Biham, E., Dunkelman, O., and Keller, N. Related-Key Boomerang and Rectangle Attacks. *Advances in Cryptology - Eurocrypt'05*, LNCS 3494, Springer, pp. 507–525 (2005).
5. Borisov, N., Chew, M., Johnson, R., and Wagner, D. Multiplicative Differentials. *Proc. of FSE'02*, LNCS 2365, Springer, pp. 17–33 (2002).
6. Boyd, C., and Mathuria, A. *Protocols for Authentication and Key Establishment*. Springer (2003).
7. Canetti, R. Universally Composable Security: A New Paradigm for Cryptographic Protocols. *Proc. of IEEE-FOCS'01*, pp. 136–145 (2001).
8. Diffie, W., and Hellman, M. E. Privacy and Authentication: An Introduction to Cryptography. *Proc. of the IEEE* 67, 3, pp. 397–427 (1979).
9. Ferguson, N., Kelsey, J., Lucks, S., Schneier, B., Stay, M., Wagner, D., and Whiting, D. Improved Cryptanalysis of Rijndael. *Proc. of FSE'00*, LNCS 1978, Springer, pp. 213–230 (2000).
10. Gilbert, H., and Minier, M. A Collision Attack on 7 Rounds of Rijndael. *Proc. of AES'00*, pp. 230–241 (2000).
11. Goldreich, O., and Krawczyk, H. On the Composition of Zero-Knowledge Proof Systems. *Proc. of ICALP'90*, LNCS 443, Springer, pp. 268–282 (1990).
12. Iwata, T., and Kurosawa, K. OMAC: One-Key CBC MAC. *Proc. of FSE'03*, LNCS 2887, Springer, pp. 129–153 (2003).
13. Jaulmes, É., Joux, A., and Valette, F. RMAC: A Randomized MAC Beyond the Birthday Paradox Limit. Available online at <http://csrc.nist.gov/CryptoToolkit/modes/proposedmodes/rmac/rmac-spec.pdf>.
14. Kanjanarin, W., and Amornraksa, T. Scrambling and Key Distribution Scheme for Digital Television. *Proc. of IEEE-ICON'01*, pp. 140–145 (2001).
15. Kelsey, J., Schneier, B., and Wagner, D. Key-Schedule Cryptanalysis of IDEA, G-DES, GOST, SAFER, and Triple-DES. *Advances in Cryptology - Crypto'96*, LNCS 1109, Springer, pp. 237–251 (1996).
16. Kelsey, J., Schneier, B., and Wagner, D. Related-key cryptanalysis of 3-WAY, Biham-DES, CAST, DES-X, NewDES, RC2, and TEA. *Proc. of ICICS'97*, LNCS 1334, Springer, pp. 233–246 (1997).
17. Kelsey, J., Schneier, B., and Wagner, D. Protocol Interactions and the Chosen Protocol Attack. *Proc. of SPW'97*, LNCS 1361, Springer, pp. 91–104 (1998).
18. Kim, J., Biryukov, A., Preneel, B., and Hong, S. On the Security of HMAC and NMAC based on HAVAL, MD4, MD5, SHA-0 and SHA-1. *Proc. of SCN'06, to appear*, LNCS, Springer.
19. Kim, J., Biryukov, A., Preneel, B., and Lee, S. On the Security of Encryption Modes of MD4, MD5 and HAVAL. *Proc. of ICICS'05*, LNCS 3783, Springer, pp. 147–158 (2005).
20. Knudsen, L. R. Cryptanalysis of LOKI. *Advances in Cryptology - Asiacrypt'91*, LNCS 739, Springer, pp. 22–35 (1993).
21. Knudsen, L. R., and Kohno, T. Analysis of RMAC. *Proc. of FSE'03*, LNCS 2887, Springer, pp. 182–191 (2003).
22. Ko, Y., Hong, S., Lee, W., Lee, S., and Kang, J.-S. Related Key Differential Attacks on 27 Rounds of XTEA and Full-Round GOST. *Proc. of FSE'04*, LNCS 3017, Springer, pp. 299–316 (2004).

23. Kurosawa, K., and Iwata, T. TMAC: Two-Key CBC MAC. *Topics in Cryptology - CT-RSA '03*, LNCS 2612, Springer, pp. 33–49 (2003).
24. Lucks, S. Ciphers Secure against Related-Key Attacks. *Proc. of FSE'04*, LNCS 3017, Springer, pp. 359–370 (2004).
25. NIST. Advanced Encryption Standard (AES). *FIPS 197* (2001).
26. NIST. Recommendation for Block Cipher Modes of Operation. *SP800-38A* (2001).
27. NIST. Recommendation for Block Cipher Modes of Operation: The CCM Mode for Authentication and Confidentiality. *SP800-38C* (2004).
28. NIST. Recommendation for Block Cipher Modes of Operation: The CMAC Mode for Authentication. *SP800-38B* (2005).
29. NIST. Recommendation for Block Cipher Modes of Operation: Galois/Counter Mode (GCM) for Confidentiality and Authentication. *Draft SP800-38D* (2006).
30. Pfitzmann, B., and Waidner, M. Composition and Integrity Preservation of Secure Reactive Systems. *Proc. of ACM-CCS'00*, pp. 245–254 (2000).
31. Phan, R. C.-W. Related-Key Attacks on Triple-DES and DESX Variants. *Topics in Cryptology - CT-RSA '04*, LNCS 2964, Springer, pp. 15–24 (2004).
32. Phan, R. C.-W., and Handschuh, H. On Related-Key and Collision Attacks: The Case for the IBM 4758 Cryptoprocessor. *Proc. of ISC'04*, LNCS 3225, Springer, pp. 111–122 (2004).
33. Preneel, B. Hash Functions and MAC Algorithms Based on Block Ciphers. *Proc. of IMA '97*, LNCS 1355, Springer, pp. 270–282 (1997).
34. Preneel, B., Govaerts, R., and Vandewalle, J. Differential Cryptanalysis of Hash Functions Based on Block Ciphers. *Proc. of ACM-CCS'93*, pp. 183–188 (1993).
35. Preneel, B., Govaerts, R., and Vandewalle, J. Hash Functions Based on Block Ciphers: A Synthetic Approach. *Advances in Cryptology - Crypto'93*, LNCS 773, Springer, pp. 368–378 (1994).
36. Rabin, M. O. Digitalized Signatures. *Fdns. of Secure Computation*, pp. 155–166 (1978).
37. Tsudik, G., and Herreweghen, E. V. On Simple and Secure Key Distribution. *Proc. of ACM-CCS'93*, pp. 49–57 (1993).
38. Winternitz, R. S., and Hellman, M. E. Chosen-key Attacks on a Block Cipher. *Cryptologia* 11, 1, pp. 16–20 (1987).

New Key Generation Algorithms for the XTR Cryptosystem^{*}

Maciej Grześkowiak

Adam Mickiewicz University,
Faculty of Mathematics and Computer Science,
Umultowska 87, 61-614 Poznań, Poland
maciejg@amu.edu.pl

Abstract. A. K. Lenstra and E. R. Verheul introduced a new public key system called XTR. They proposed two algorithms for finding primes p and q , where $q|(p^2 - p + 1), p \equiv 2 \pmod{3}$, which are the key parameters for the XTR. One is unable to estimate in a simple way a running time the above algorithms, nor to give a mathematical proof of their correctness or prove that these algorithms works in polynomial time as suggested authors above mentioned cryptosystem. In this paper we propose theoretical algorithms which find primes as above. We give a mathematical proof of its correctness, under the assumption of some conjecture.

1 Introduction

Many new cryptosystems have been proposed in recent years, which require generating primes of special forms as key parameters [4], [5], [6], [10], [14], [16]. For instance one is interested in generating large primes p and q , such that q divides $\Phi_n(p)$, where $\Phi_n(X)$ are n th cyclotomic polynomials for a fixed positive integer n . In some cases it is essential to find a prime p as above which is congruent to a modulo b for fixed positive integers a, b .

For $n = 6$, $a = 2$ and $b = 3$ two straightforward algorithms were proposed in [10] which find primes p and q as above, but analysis of computational complexity of these was not given there. The first one randomly chooses a positive integer r_1 until $q = \Phi_6(r_1)$ is not a prime and next finds a prime p , such that $p \equiv r_1 \pmod{q}$ and $p \equiv 2 \pmod{3}$. The second algorithm select randomly a prime $q \equiv 7 \pmod{12}$ and computes r_i for $i = 2, 3$ roots of $\Phi_6(X) \pmod{q}$. Next the algorithm select a prime p such that $p \equiv r_i \pmod{q}$ and $p \equiv 2 \pmod{3}$ for $i = 2$ or $i = 3$.

One is unable to estimate in a simple way the running time the above algorithms, nor to give a mathematical proof of their correctness or prove that these algorithms work in polynomial time. We denote by $P(a, q)$ the least prime

^{*} The work described in this paper has been supported in part by a Foundation for Polish Science scholarship and by the State Committee for Scientific Research under Grant 0T00A00323.

$p \equiv a \pmod{q}$ for any $a, q \in \mathbf{N}, 1 \leq a \leq q, (a, q) = 1$. A well known theorem by Heath-Brown [8] gives $P(a, q) \ll q^{5.5}$. From this we cannot exclude possibility that the number of steps we need to find prime p is exponential. On the other hand we don't know if there exist infinitely many primes of the form $\Phi_6(r_1)$. This is an extremely hard mathematical problem still unproved up to now. On the other hand it is worth pointing out that the above algorithms work perfectly well in practise. Analysis and theoretical estimation of computational complexity of algorithms from [10] under the assumption of some unproved conjectures can be found in [7].

In this paper we take a theoretical approach to generate primes as above. Our methods don't require to generate a prime in an arithmetic progression in order to exclude theoretical possibility exponential the number of steps of the algorithm. An analysis of running time of our algorithms is possible only under the assumption of hypotheses related to primes which are values of irreducible polynomials in one variable with integer coefficients and positive leading coefficient.

2 Hypothesis H

In 1958 Schinzel and Sierpiński proposed a conjecture concerning prime values of irreducible polynomials in one variable with integer coefficients and positive leading coefficient [15]. It is known as Hypothesis H. P. T Bateman and R. A. Horn [2] gave a quantitative form of it.

Hypothesis 1 (H). *Suppose $f_1, f_2, \dots, f_k \in \mathbf{Z}[x]$ are polynomials with positive leading coefficients of degrees h_1, h_2, \dots, h_k respectively, and such that $f(n) = f_1(n)f_2(n) \dots f_k(n)$ has no fixed divisor greater than 1. Suppose that polynomials are irreducible over the field of rational numbers and no two of them are identical. Let $\mathcal{Q}_N(f_1, f_2, \dots, f_k)$ denote the number of positive integers n between 1 and N such that all numbers $f_1(n), f_2(n), \dots, f_k(n)$ are prime. Then as $N \rightarrow \infty$ we have*

$$\mathcal{Q}_N(f_1, \dots, f_k) = \frac{C(f_1, \dots, f_k)}{h_1 \dots h_k} \int_2^N \frac{du}{\log^k u} + o\left(\int_2^N \frac{du}{\log^k u}\right), \tag{1}$$

where

$$C(f_1, \dots, f_k) = \prod_p \left\{ \left(1 - \frac{1}{p}\right)^{-k} \left(1 - \frac{\rho(p)}{p}\right) \right\},$$

the product being taken over all primes and $\rho(p)$ being the number of solutions of the congruence

$$f_1(n) \dots f_k(n) \equiv 0 \pmod{p}.$$

Note that when $\rho(p) < p$ for all p , the constant $C(f_1, \dots, f_k)$ is positive [3] and so (1) can then be written as

$$Q_N(f_1, \dots, f_k) \sim \frac{C(f_1, \dots, f_k)}{h_1 \dots h_k} \int_2^N \frac{du}{\log^k u}. \tag{2}$$

3 The Main Algorithm

In this section we present an algorithm for generating key parameters for the XTR public key system. For the convenience of this analysis, algorithms are written in particular pseudocode. First we introduce the meaning of functions which will be used in our algorithms.

The function *Random*(r) returns a random integer r which fulfils conditions described after ‘.’. Let n_1, n_2, \dots, n_r be positive integers. The function *IsPrimeQuick*(n_1, n_2, \dots, n_r) returns **true** if every n_1, n_2, \dots, n_r satisfies probabilistic test for primality or some primality test which is deterministic under the assumption some of hypotheses and both run in polynomial time. Otherwise function *IsPrimeQuick*(n_1, n_2, \dots, n_r) returns **false**. Examples of such algorithms can be found in [12], [13].

The function *IsPrime*(n_1, n_2, \dots, n_r) returns **true** if every n_1, n_2, \dots, n_r satisfies some deterministic primality test which runs in polynomial time, and **false** otherwise. For this purpose one can use Agrawal-Kayal-Saxena primality test [1]. In original version it runs in deterministic $O(\log^{12} n)$ time when testing an integer n . However, the current best unconditional bound of AKS test has been reduced to $O(\log^6 n)$, see [11].

By \mathcal{PT} we will denote the number of bit operations necessary to carry out deterministic primality test. For simplicity assume that $\mathcal{PT} \gg \log^3 n$.

We need the following lemma

Lemma 1. *Let*

$$\begin{aligned} Q_1(S, X) &= (27S^2 + 36S + 12)X^2 - (9S^2 + 6S)X + 3S^2 + 3S + 1, \\ Q_2(S, X) &= (27S^2 + 36S + 12)X^2 + (9S^2 + 6S)X + 3S^2 + 3S + 1, \\ Q_3(S, X) &= (27S^2 + 18S + 3)X^2 - (9S^2 + 12S + 3)X + 3S^2 + 3S + 1, \\ Q_4(S, X) &= (27S^2 + 18S + 3)X^2 + (9S^2 + 12S + 3)X + 3S^2 + 3S + 1, \\ P_1(S, X) &= (27S + 18)X^2 - (9S + 3)X + 3S + 2, \\ P_2(S, X) &= (27S + 18)X^2 + (9S + 3)X + 3S + 2, \\ P_3(S, X) &= (27S + 9)X^2 - (9S + 6)X + 3S + 2, \\ P_4(S, X) &= (27S + 9)X^2 + (9S + 6)X + 3S + 2, \\ Q_5(X) = Q_7(X) &= 9X^2 - 3X + 1, \\ Q_6(X) = Q_8(X) &= 9X^2 + 3X + 1, \\ F(X) &= X^2 - X + 1 \end{aligned} \tag{3}$$

are polynomials in $\mathbf{Z}[S, X]$. Then

$$F(P_j(S, X)) = 3Q_j(S, X)Q_{j+4}(X) \tag{4}$$

for $j = 1, \dots, 4$. Moreover for every fixed positive integer s polynomials $Q_j(s, X)$, $Q_{j+4}(X)$, $P_j(s, X) \in \mathbf{Z}[X]$ are irreducible in $\mathbf{Q}[X]$ for $j = 1, \dots, 4$.

Proof. A trivial verification shows that

$$F(P_j(S, X)) - 3Q_j(S, X)Q_{j+4}(X) = 0$$

for $j = 1, \dots, 4$. Note that discriminants $D(Q_{j+4})$ of polynomials $Q_{j+4}(X)$ are less than 0 for $j = 1, \dots, 4$. Moreover, let s be a fixed positive integer, then an easy computation shows that

$$\begin{aligned} D(Q_j) &= -3(3s + 2)^4 && \text{for } j = 1, 2, \\ D(Q_j) &= -3(3s + 1)^4 && \text{for } j = 3, 4, \\ D(P_j) &= -27(s + 1)(9s + 5) && \text{for } j = 1, 2, \\ D(P_j) &= -27(3s + 2)(9s + 2) && \text{for } j = 3, 4. \end{aligned}$$

Since $\deg(Q_{j+4}(X)) = \deg(Q_j(s, X)) = \deg(P_j(s, X)) = 2$ for $j = 1, \dots, 4$, polynomials are irreducible in $\mathbf{Q}[X]$. □

Now we present the algorithm which generates primes p and q such that $q|p^2 - p + 1$, where $p \equiv 2 \pmod{3}$. Let us fix a pair of polynomials P_j, Q_i from Lemma 1, where $j = 1, \dots, 4$ and $i = j$ or $i = j + 4$ and an arbitrary positive integer A .

Algorithm 2. (selection of $p \equiv 2 \pmod{3}$ and q , such that $q|p^2 - p + 1$)

1. *Random*(s) : $s \in [1, A]$, $s \in \mathbf{N}$;
2. do
3. do
4. *Random*(r) : $r \in [N, 2N]$, $r \in \mathbf{N}$;
5. $p := P_j(s, r)$;
6. $q := Q_i(s, r)$; ¹
7. while not *IsPrimeQuick*(p, q) ;
8. while not *IsPrime*(p, q) ;
9. return (p, q) .

Theorem 3. Assume Hypothesis H. Then there exist constants b_0 and N_0 such that for every integer $N \geq N_0$ and an arbitrary real $\lambda \geq 1$, Algorithm 2 generates primes $p \equiv 2 \pmod{3}$ and q , key parameters for the XTR cryptosystem, such that $q|p^2 - p + 1$ and $q \asymp N^2$, $p \asymp N^2$ with probability greater than or equal to $1 - e^{-\lambda}$ after repeating $\lceil b_0 \lambda \log^2 N \rceil$ steps of the algorithm. Every step of the algorithm takes no more than \mathcal{PT} bit operations.

In order to prove Theorem 3 we need the following technical lemma.

Lemma 2. Let s be a fixed natural number and

$$H_j(s, X) = P_j(s, X)Q_j(s, X)Q_{j+4}(X) \in \mathbf{Z}[X],$$

where polynomials P_j, Q_j, Q_{j+4} are (3) for $j = 1, \dots, 4$. Then, $H_j(s, x)$ has no fixed divisor > 1 .

¹ Note that polynomials Q_i depends only on r for $i = 5, \dots, 8$.

Proof. We first observe that for any fixed integer s

$$\begin{aligned} H_1(s, 0) &= H_2(s, 0) = (3s + 2)(3s^2 + 3s + 1), \\ H_1(s, 1) &= H_2(s, -1) = 7(21s + 17)(21s^2 + 33s + 13) \\ H_1(s, -1) &= H_2(s, 1) = 13(39s + 23)(39s^2 + 45s + 13). \end{aligned}$$

Similarly

$$\begin{aligned} H_3(s, 0) &= H_4(s, 0) = (3s + 2)(3s^2 + 3s + 1), \\ H_3(s, -1) &= H_4(s, 1) = 13(39s + 17)(39s^2 + 33s + 7), \\ H_3(s, -4) &= H_4(s, 4) = 157(471s + 170)(471s^2 + 339s + 61). \end{aligned}$$

We show that

$$\begin{aligned} (H_1(s, 0), H_1(s, 1), H_1(s, -1)) &= 1, \\ (H_3(s, 0), H_3(s, -1), H_3(s, -4)) &= 1. \end{aligned} \tag{5}$$

We give the proof only for the first case. The second case is similar and we left it to the reader. Suppose that there is a prime p such that $p \mid 3s + 2$ and $p \mid 21s + 17$. Then $p \mid 21s + 14$. Since p divides difference last two numbers then $p \mid 3$. Since $3 \nmid 3s + 2$ and $3 \nmid 21s + 17$ we get

$$(3s + 2, 21s + 17) = 1 \quad \text{for } s \in \mathbf{N}. \tag{6}$$

Now we suppose that there is a prime p such that $p \mid 3s + 2$ and $p \mid 21s^2 + 33s + 13$. Then $p \mid 6(3s + 2)^2$ and $p \mid 3(21s^2 + 33s + 13)$ and so $p \mid 15s + 11$. By assumption $p \mid 3(3s + 2)$ this gives $p \mid 1$. We obtain that

$$(3s + 2, 21s^2 + 33s + 13) = 1 \quad \text{for } s \in \mathbf{N}. \tag{7}$$

We observe that $7 \mid 3s + 2$ when $s \equiv 4 \pmod{7}$, but in this case $7 \nmid 39s + 23$ and $7 \nmid 39s^2 + 45s + 13$. Moreover $13 \mid 3s + 2$ for $s \equiv 8 \pmod{13}$, but then $13 \nmid 21s + 17$ and $13 \nmid 21s^2 + 33s + 13$. From (6), (7) we have that

$$(3s + 2, H_1(s, 1), H_1(s, -1)) = 1 \quad \text{for } s \in \mathbf{N}. \tag{8}$$

Suppose that there is a prime p such that $p \mid 3s^2 + 3s + 1$ and $p \mid 21s^2 + 33s + 13$. Then $p \mid 12s + 6$ which gives that $p \mid 2$ or $p \mid 3$ or $p \mid 2s + 1$. In the last case $p \mid 3(2s + 1)^2$ and by assumptions $p \mid 4(3s^2 + 3s + 1)$ we have $p \mid 1$. Because $3s^2 + 3s + 1 \equiv 1 \pmod{2, 3}$ and $21s^2 + 33s + 13 \equiv 1 \pmod{2, 3}$, we have

$$(3s^2 + 3s + 1, 21s^2 + 33s + 13) = 1 \quad \text{for } s \in \mathbf{N}. \tag{9}$$

Now we suppose that there is a prime p such that $p \mid 3s^2 + 3s + 1$ and $p \mid 21s + 17$. Then $p \mid (21s + 17)^2$ and $p \mid 147(3s^2 + 3s + 1)$ which gives $p \mid 273s + 142$. Moreover, $p \mid 13(21s + 17)$ and consequently $p \mid 79$. It is easy to see that $79 \mid 21s + 17$ and

$79 \mid 3s^2 + 3s + 1$ when $s \equiv 18 \pmod{79}$, but in this case $79 \nmid 39s + 23$ and $79 \nmid 39s^2 + 45s + 13$. We conclude that

$$(21s + 17, 3s^2 + 3s + 1, H_1(s, -1)) = 1 \quad \text{for } s \in \mathbf{N}. \tag{10}$$

We observe that $7 \mid 3s^2 + 3s + 1$ when $s \equiv 1, 5 \pmod{7}$, but then $7 \nmid 39s + 23$ and $7 \nmid 39s^2 + 45s + 13$. Moreover we can see that $13 \mid 3s^2 + 3s + 1$ when $s \equiv 5, 7 \pmod{13}$, but then $13 \nmid 21s + 17$ and $13 \nmid 21s^2 + 33s + 13$. We conclude from (9), (10) that

$$(3s^2 + 3s + 1, H_1(s, 1), H_1(s, -1)) = 1 \quad \text{for } s \in \mathbf{N}. \tag{11}$$

Finally (5) follows from (8),(11). □

Now we give the proof of Theorem 3.

Proof. Let s be an integer, $s \in [1, A]$, where A denotes an arbitrary fixed constant. Let $r \in [N, 2N], r \in \mathbf{N}$. Assume that the algorithm finds primes $p = P_j(s, r)$ and $q = Q_i(s, r)$, for fixed $j = 1, \dots, 4$ and $i = j$ or $i = j + 4$. It follows from Lemma 1 that $q \mid p^2 - p + 1, p \equiv 2 \pmod{3}$, so that the algorithm generates key parameters for the XTR cryptosystem.

By Lemma 1 polynomials $P_j(s, X), Q_i(s, X)$ are irreducible in $\mathbf{Q}[X]$. Let $H_j(s, X) = P_j(s, X)Q_i(s, X) \in \mathbf{Q}[X]$. Lemma 2 implies that there not exists an integer $m > 1$ divides $H_j(s, x)$ for every integer x . Therefore polynomials $P_j(s, X), Q_i(s, X)$ satisfies assumptions of Hypothesis H. We conclude from (1) that there exists an integer N_0 such that for every integer $N \geq N_0$

$$\begin{aligned} \mathcal{Q}_N(P_j(s), Q_i(s)) &= |\{r : 1 \leq r \leq N : P_j(s, r) = p, Q_i(s, r) = q\}| \\ &= \frac{C(s)}{4} \int_2^N \frac{du}{\log^2 u} + o\left(\int_2^N \frac{du}{\log^2 u}\right), \end{aligned}$$

where p, q are prime and

$$C(s) = \prod_l \left\{ \left(1 - \frac{1}{l}\right)^{-2} \left(1 - \frac{\rho(l)}{l}\right) \right\}, \tag{12}$$

the product (12) being taken over all primes and $\rho(l)$ being the number of solutions of the congruence $H_j(s, X) \equiv 0 \pmod{l}$. By (2) we obtain

$$|\{r : N \leq r \leq 2N, P_j(s, r) = p, Q_i(s, r) = q\}| \sim C_0(s) \frac{N}{\log^2 N},$$

where $C_0(s) = C(s)4^{-1}$. We denote by A_r the event that a random integer $r \in [N, 2N]$ is such that $P_j(s, r)$ and $Q_j(s, r)$ are prime. Hence the probability that in k trials A_r does not occur is

$$\begin{aligned} \left(1 - \frac{C_0(s)}{\log^2 N}\right)^k &= \exp\left(k \log\left(1 - \frac{C_0(s)}{\log^2 N}\right)\right) \leq \\ &\leq \exp\left(\frac{-C_0(s)k}{\log^2 N}\right) \leq e^{-\lambda} \end{aligned}$$

for an arbitrary real $\lambda \geq 1$ and $k = b_0 \lambda \log^2 N$, where $b_0 = C_0(s)^{-1}$. Hence the probability that in k trials A_r does occur is greater or equal to $1 - e^{-\lambda}$. So after repeating $[b_0 \lambda \log^2 N]$ steps, the algorithm finds an integer r and primes $p = P_j(s, r)$ and $q = Q_j(s, r)$ with probability greater or equal to $1 - e^{-\lambda}$.

Now, we estimate the number of bit operations required to carry out the steps of the algorithm. It takes a fixed numbers of time to generate a random bit, and $O(\log N)$ bit operations to generate a random integer $r \in [N, 2N]$. Polynomials' P_j, Q_i value computation can be done with $O(\log^2 N)$ bit operations. Probabilistic primality test takes no more than $O(\log^3 N)$ operations. The most time-consuming step of the algorithm is the deterministic primality test for numbers p and q which takes no more than \mathcal{PT} operations. \square

3.1 General Version of Algorithm 2

The main idea of this algorithm is similar to method from section 3, but analysis of running-time is more complicated. Let us fix an arbitrary positive integer A .

Algorithm 4. (*selection of $p \equiv 2 \pmod{3}$ and q , such that $q | p^2 - p + 1$*)

```

1. success:= false ;
2. Random(s) :  $s \in [1, A], s \not\equiv 1 \pmod{7}, s \in \mathbf{N}$  ;2
3. do
4.     do
5.         Random(r) :  $r \in [N, 2N], r \in \mathbf{N}$  ;
6.         for j := 1 to 4 do
7.             for i := 0 to 1 do {
8.                  $p := P_j(s, r)$  ;
9.                  $q := Q_{j+4i}(s, r)$  ;3
10.                if IsPrimeQuick(p, q)
11.                    then success := true ;
12.                    break ; }4
13.         while not success ;
14.     while not IsPrime(p, q) ;
15.     return (p, q) .

```

Theorem 5. *Assume Hypothesis H for any pair of polynomials $(P_j(X, s), Q_i(X, s))$ (see 3), where $j = 1, \dots, 4$ and $i = j$ or $i = j + 4$. Then there exist constants b_1 and N_1 such that for every integer $N \geq N_1$ and an arbitrary real $\lambda \geq 1$, Algorithm 4 generates primes $p \equiv 2 \pmod{3}$ and q , key parameters for the XTR cryptosystem, such that $q | p^2 - p + 1$ and $q \asymp N^2, p \asymp N^2$ with probability greater than or equal to $1 - e^{-\lambda}$ after repeating $[b_1 \lambda \log^2 N]$ steps of the algorithm. Every step of the algorithm takes no more than \mathcal{PT} bit operations.*

In order to prove Theorem 5 we need the following technical lemma

² The assumption that $s \not\equiv 1 \pmod{7}$ we need only for analysis of this algorithm. In practice we can choose any $s \in \mathbf{N}$.

³ Note that for $i = 1$ polynomial Q_{j+4i} depends only on r .

⁴ Go to 14.

Lemma 3. *Let $s \not\equiv 1 \pmod{7}$ be a fixed natural number and*

$$H(s, X) = \prod_{j=1}^4 P_j(s, X) Q_j(s, X) \prod_{k=5}^6 Q_k(X) \in \mathbf{Z}[X].$$

Moreover let $\rho(p)$ denote the number of solutions of the congruence $H(s, X) \equiv 0 \pmod{p}$. Then $\rho(p) < p$ for all primes p .

Proof. Factors of $H(s, X) \in \mathbf{Z}[X]$ are irreducible in $\mathbf{Q}[X]$ polynomials of degree 2 and coefficients of P_j, Q_j for $j = 1, \dots, 4$ and Q_k for $k = 5, 6$ are coprime for every $s \in \mathbf{N}$. Hence factors of $H(s, X)$ can have no more than two roots modulo p , where p is prime. Then $\rho(p) \leq 20$ for $p > 20$. The proof for $p < 20$ is a matter of straightforward computation. We leave it to the reader.

Now we give the proof of Theorem 5

Proof. Let $s \not\equiv 1 \pmod{7}$ be an integer, $s \in [1, A]$, where A denotes an arbitrary fixed constant. It follows from Lemma 1 that the algorithm generates key parameters for the XTR cryptosystem. Now, we denote by A_i the event that a random integer $r \in [N, 2N]$ is such that at least one pair of polynomials

$$(P_j(s, r), Q_i(s, r)) \text{ for } j = 1, \dots, 4 \text{ and } i = j \text{ or } i = j + 4 \tag{13}$$

has simultaneously prime values. Then the probability that at least one A_i occurs is

$$\begin{aligned} P\left(\bigcup_{i=1}^8 A_i\right) &= \sum_{i=1}^8 P(A_i) - \sum_{1 \leq i_1 < i_2 \leq 8} P(A_{i_1} \cap A_{i_2}) + \dots \\ &\dots + \sum_{1 \leq i_1 < \dots < i_7 \leq 8} P(A_{i_1} \cap \dots \cap A_{i_7}) - P(A_{i_1} \cap \dots \cap A_{i_8}). \end{aligned} \tag{14}$$

As a first step we will bound the first term in (14). Let $H_i(s, X) = P_j(s, X) Q_i(s, X) \in \mathbf{Q}[X]$ for $j = 1, \dots, 4$ and $i = j$ or $i = j + 4$.

We conclude from Lemmas 1 and 2 that every pair of (13) satisfies assumptions of Hypothesis H. Then there exists an integer N_1 such that for every integer $N \geq N_1$

$$\begin{aligned} \Omega_N(P_j(s), Q_i(s)) &= |\{r : 1 \leq r \leq N : P_j(s, r) = p, Q_i(s, r) = q\}| \\ &= \frac{C_i(s)}{4} \int_2^N \frac{du}{\log^2 u} + o\left(\int_2^N \frac{du}{\log^2 u}\right), \end{aligned}$$

where p, q are prime and

$$C_i(s) = \prod_l \left\{ \left(1 - \frac{1}{l}\right)^{-2} \left(1 - \frac{\rho(l)}{l}\right) \right\}, \tag{15}$$

the product (15) being taken over all primes and $\rho_i(l)$ being the number of solutions of the congruence $H_i(s, X) \equiv 0 \pmod{l}$, where $j = 1, \dots, 4$, and $i = j$ or $i = j + 4$. Hence we obtain that

$$|\{r : N \leq r \leq 2N, P_j(s, r) = p, Q_i(s, r) = q\}| = C_i^0(s) \frac{N}{\log^2 N} + o\left(\frac{N}{\log^2 N}\right),$$

where $C_i^0(s) = C_i(s)4^{-1}$, $i = 1, \dots, 8$. Then

$$\sum_{i=1}^8 P(A_i) = \frac{c_1(s)}{\log^2 N} + o\left(\frac{1}{\log^2 N}\right), \tag{16}$$

where

$$c_1(s) = \sum_{i=1, \dots, 8} C_i^0(s).$$

To estimate other terms of (14) we use the theorem below

Theorem 6. *Let $F_1(n), \dots, F_g(n)$ be distinct irreducible polynomials with integral coefficients, and positive leading coefficient. Write*

$$F(n) = F_1(n) \cdots F_g(n),$$

let $\rho(p)$ denote the number of solutions of

$$F(n) \equiv 0 \pmod{p},$$

and suppose that

$$\rho(p) < p, \text{ for all primes } p.$$

Let x and y be real numbers satisfying

$$1 < y \leq x.$$

Then

$$\begin{aligned} & |\{n : x - y < n \leq x, F_i(n) \text{ prime for } i = 1, \dots, g\}| \\ & \leq 2^g g! \prod_p \left(1 - \frac{\rho(p)}{p}\right) \left(1 - \frac{1}{p}\right)^{-g} \frac{y}{\log^g y} \left\{1 + O_F\left(\frac{\log \log 3y}{\log y}\right)\right\}. \end{aligned}$$

Proof. see [9] □

It is easily seen that

$$\begin{aligned} & A_j \cap A_{j+4} \text{ for } j = 1, \dots, 4 \\ \text{and } & A_k \cap A_{k+2}, \text{ for } k = 5, 6 \end{aligned} \tag{17}$$

describe events that a random integer $r \in [N, 2N]$ is such that the corresponding three polynomials attain prime values and similarly for intersection (14) with more terms. We conclude from Lemmas 1 and 3 and Theorem 6 that

$$\sum_{1 \leq i_1 < i_2 \leq 8} P(A_{i_1} \cap A_{i_2}) \leq \frac{c_2(s)}{\log^3 N} + O_s \left(\frac{\log \log N}{\log^4 N} \right), \tag{18}$$

where

$$c_2(s) = \sum_{k=1, \dots, 6} c_k(s)$$

and

$$c_k(s) = 2^3 3! \prod_p \left(1 - \frac{\rho_k(p)}{p} \right) \left(1 - \frac{1}{p} \right)^{-3}$$

the product (15) being taken over all primes and $\rho_k(p)$ being the number of solutions of the congruence $G_k(s, X) \equiv 0 \pmod{p}$, where $G_k(s, X) \in \mathbf{Z}[X]$ denotes the product of irreducible three polynomials corresponding to events (17) for $k = 1, \dots, 6$. Then from (16) and (18), we have that

$$P \left(\bigcup_{i=1}^8 A_i \right) \geq \frac{c_1(s)}{2 \log^2 N} = \frac{c_2(s)}{\log^2 N},$$

where $c_2(s) = \frac{c_1(s)}{2}$. Hence the probability that in k trials every $A_i, i = 1, \dots, 8$ does not occur is less or equal to

$$\begin{aligned} \left(1 - \frac{c_2(s)}{\log^2 N} \right)^k &= \exp \left(k \log \left(1 - \frac{c_2(s)}{\log^2 N} \right) \right) \leq \\ &\leq \exp \left(\frac{-c_2(s)k}{\log^2 N} \right) \leq e^{-\lambda} \end{aligned}$$

for an arbitrary real $\lambda \geq 1$ and $k = b_1 \lambda \log^2 N$, where $b_1 = c_2(s)^{-1}$. We conclude that the probability that in k trials at least one of $A_i, i = 1, \dots, 8$ does occur is greater or equal to $1 - e^{-\lambda}$. From the above it follows that after repeating $[b_1 \lambda \log^2 N]$ steps, the algorithm finds an integer r such that at least one pair of (13) which have prime values with probability greater or equal to $1 - e^{-\lambda}$. The number of bit operations required to carry out the steps of the algorithm is similar to the one used in Algorithm 2. □

References

1. Agrawal M., Kayal K., Saxena N.: *Primes is P*, Ann. of Math, 160, 2004, 781-793.
2. Bateman P. T., Horn R. A.: *A heuristic asymptotic formula concerning the distribution of prime numbers*, Math. Comp, 16, 1962, 119-132.

3. Bateman P. T., Horn R. A.: *Primes represented by irreducible polynomials in one variable*, Proc. Symposia in Pure Mathematics, 8, 1965, 119-132.
4. Gong G., Harn L.: *Public-key cryptosystems based on cubic finite field extension*, IEEE IT, 45, 7, 1999, 2601-2605.
5. Giuliani K., Gong G.: *Analogues to the Gong-Harn and XTR cryptosystem*, Combinatorics and Optimization Research Report CORR 2003-34, University of Waterloo, 2003.
6. Giuliani K., Gong G.: *Efficient key agreement and signature scheme using compact representation in $GF(p^{10})$* , Proceedings of the 2004 IEEE International Symposium on Information Theory, Chicago, 2004, 13-30.
7. Grześkowiak M.: *Analysis of algorithms of generating key parameters for the XTR cryptosystem*, Proceedings of WARTACRYPT'2004. Tatra Mountains Mathematical Publications, 35, 2006.
8. Heath-Brown D. R.: *Zero-free regions for Dirichlet L-functions and the least prime in an arithmetic progression*, Proc. London Math. Soc., 64, 3, 1992, 265-338.
9. Halberstam H., Richert H. E.: *Sieve methods*, Academic Press, London 1974.
10. Lenstra A. K., Verheul E. R.: *The XTR public key system*, Proceedings of Crypto 2000, LNCS 1880, Springer-Verlag, 2000, 1-19.
11. Lenstra H. W. Jr., Pomerance C.: *Primality testing with Gaussian periods*, Manuscript. March 2003.
12. Miller G. L.: *Riemann's hypothesis and tests for primality*, J. Comput. Systems Sci., 13, 1976, 300-317.
13. Rabin M. O.: *Probabilistic algorithm for testing primality*, J. Number Theory, 12, 1980, 128-138.
14. Rubin K, Silverberg.: *Using primitive subgroups to do more with fewer bits*, ANTS-VI, LNCS 3076, 18-41, 2004.
15. Schinzel A., Sierpiński W.: *Sur certaines hypothèses concernant les nombres premiers*, Acta Arith., 4, 1956, 185-208.
16. Smith P., Skinner.: *A public key cryptosystem and a digital signature system based on the Lucas function Analogue to Discrete Logarithms*, Advances in Cryptology Asiacypt'94, LNCS 917, Springer-Verlag, 1994, 357-364.

Public-Key Encryption from ID-Based Encryption Without One-Time Signature

Chik How Tan

NISlab, Department of Computer Science and Media Technology
Gjøvik University College, Norway
chik.tan@hig.no

Abstract. Design a secure public key encryption scheme and its security proof are one of the main interests in cryptography. In 2004, Canetti, Halevi and Katz [8] constructed a public key encryption (PKE) from a selective identity-based encryption scheme with a strong one-time signature scheme. In 2005, Boneh and Katz [6] improved Canetti-Halevi-Katz construction by replacing a strong one-time signature with a message authentication code, but it is not publicly verifiable. Later, Boyen, Mei and Waters [7] constructed PKE scheme directly from Waters' IBE scheme [17], which is only secure against direct chosen-ciphertext attack and is not secure against adaptive chosen-ciphertext attack. In 2006, Tan [16] further improved the efficiency of Canetti-Halevi-Katz (CHK) construction by directly from Boneh-Boyen identity based encryption (IBE) scheme [4] with a weak one-time signature. In this paper, we construct an efficient public key encryption scheme without one-time signature, which preserves a publicly verifiable property and secure against adaptive chosen-ciphertext attack. The construction of the proposed scheme is based on Boneh-Boyen identity-based encryption (IBE) scheme [2] and a trapdoor function. We also show that the proposed scheme is more efficient than CHK construction.

Keywords: Cryptography, public key encryption, bilinear maps.

1 Introduction

Design a secure public key encryption scheme and its security proof are one of the main interests in cryptography. The security notion of encryption scheme against adaptive chosen ciphertext attack (CCA2) was first introduced by Rackoff and Simon [14]. This security notion was later widely accepted to provide the right level of security for public key encryption (PKE) scheme, which is also referred as an IND-CCA2 secure scheme. In 1998, Cramer and Shoup [9] constructed the first practical and provably secure public key encryption scheme against adaptive chosen ciphertext attack under the standard assumptions in the standard model. From 1998 to 2004, Cramer-Shoup scheme and its variants remained the only practical and secured schemes in the standard model.

In 2004, Canetti, Halevi and Katz [8] gave a generic construction of public key encryption scheme, which is different from Cramer-Shoup scheme. Their con-

struction is a black-box transformation from a secure selective identity based encryption scheme against chosen-plaintext attack to an IND-CCA2 public key encryption scheme (PKE) by using strong one-time signature scheme. The main advantage of Canetti-Halevi-Katz construction (CHK) over Cramer-Shoup scheme [9] is that the validity of a ciphertext can be verified publicly; while Cramer-Shoup scheme can only be verified with a private key. Since then, few attempts try to enhance the efficiency of CHK construction. In 2005, Boneh and Katz [6] improved the efficiency by replacing a one-time signature with a message authentication code. But, the Boneh-Katz construction is no longer publicly verifiable. Later, Boyen, Mei and Waters [7] constructed a PKE scheme which is only secure against direct chosen ciphertext and is not secure against adaptive chosen ciphertext attack. In 2006, Tan [16] further improved the efficiency of CHK construction by directly from Boneh-Boyen identity based encryption (IBE) scheme [4] with a weak one-time signature. In this paper, we construct an efficient public key encryption scheme without one-time signature, which preserves the publicly verifiable property. The construction of the proposed scheme is based on Boneh-Boyen IBE scheme [2] and a trapdoor function. We showed that the proposed scheme is secure against adaptive chosen ciphertext attack and as efficient as Boneh-Boyen IBE scheme [2].

Organization of Paper

The paper is organised as follows: In Section 2, we briefly describe bilinear maps and its properties; and bilinear Diffie-Hellman inversion assumptions. In Section 3, we construct a public key encryption scheme based on Boneh-Boyen identity-based encryption (IBE) scheme and a trapdoor function. In Section 4, the computational complexity of the proposed scheme is compared with other scheme which is based on Canetti-Halevi-Katz construction. We showed that the proposed scheme is as efficient as Boneh-Boyen IBE scheme. Section 5 gives a detailed proof of the proposed encryption scheme which is secure against CCA2 under the decisional bilinear Diffie-Hellman inversion assumption (DBDHI) in the standard model.

2 Bilinear Maps and Assumptions

Let G_1 and G_2 be cyclic groups of prime order p and g be a generator of G_1 . Let e be an admissible bilinear map : $G_1 \times G_1 \rightarrow G_2$ satisfying the following:

- e is bilinear : For all $u, v \in G_1$ and $a, b \in \mathbb{Z}_p^*$, then $e(u^a, v^b) = e(u, v)^{ab}$.
- e is non-degenerate : $e(g, g) \neq 1$.
- e is efficiently computable of $e(u, v)$ for $\forall u, v \in G_1$

Definition 1. (Bilinear Diffie-Hellman Inversion Problem (BDHI)). *Given a $(q + 1)$ -tuple $(g, g^x, \dots, g^{x^q}) \in G_1^{q+1}$ where $x \in \mathbb{Z}_p^*$, output $e(g, g)^{1/x}$.*

Definition 2. (Decisional Bilinear Diffie-Hellman Inversion Problem (DBDHI)). *Given a $(q + 2)$ -tuple $(g, g^x, \dots, g^{x^q}, T) \in G_1^{q+1} \times G_2$ where $x \in \mathbb{Z}_p^*$, decide $T = e(g, g)^{1/x}$.*

Definition 3. (DBDHI Assumption). *We say that (t, ϵ) -DBDHI assumption holds if no t -time algorithm has the probability of at least ϵ in solving the DBDHI problem. That is,*

$$|\Pr[\mathcal{A}(g, g^x, \dots, g^{x^q}, e(g, g)^{1/x}) = 1] - \Pr[\mathcal{A}(g, g^x, \dots, g^{x^q}, T) = 1]| \geq \epsilon,$$

where the probability is over the random choice of $x \in Z_p^*$, the random choice of $T \in G_2$.

Now, we give a definition of a target collision resistant hash function, which was defined by Cramer and Shoup [9], as follows.

Definition 4. (Target Collision Resistance). *Let \bar{w} and \bar{n} be two positive integers. We say that a family of hash function $\mathcal{H} = \{H_k : \{0, 1\}^{\bar{w}} \rightarrow \{0, 1\}^{\bar{n}}\}_{k \in K}$ is (t, ϵ_H) -target collision resistance hash function if the probability of any t -time algorithm \mathcal{A} is*

$$\Pr[H_k(x) = H_k(y) \text{ and } y \neq x : \text{given } x \in \{0, 1\}^{\bar{w}}, k \leftarrow K; y \leftarrow \mathcal{A}(k)] < \epsilon_H.$$

3 Propose Encryption Scheme

In this section, we construct a public key encryption scheme which is called PE1. The proposed scheme is constructed from Boneh-Boyen IBE scheme (BB-E) [2] based on DBDHI assumption; and a trapdoor function. This construction converts a secure selective identity-based encryption scheme against chosen plaintext attack to a secure public key encryption against adaptive chosen ciphertext attack (IND-CCA2). The proposed scheme does not require a strong one-time signature scheme and is different from Canetti-Halevi-Katz construction. Before describing the proposed public key encryption scheme, we first give a definition of a public key encryption scheme.

A public key encryption scheme PE consists of four algorithms $(\mathcal{P}, \mathcal{K}, \mathcal{E}, \mathcal{D})$. The parameter set up \mathcal{P} selects an appropriate security parameters. The key generation algorithm \mathcal{K} generates a key pair $(pk, sk) \leftarrow \mathcal{K}$ where pk is a public key and sk is a private key. The encryption algorithm \mathcal{E} takes a public key pk and a plaintext m , returns a ciphertext $c \leftarrow \mathcal{E}(pk, m)$. The decryption algorithm \mathcal{D} takes a private key sk and a ciphertext c , returns $m = \mathcal{D}(sk, c)$.

Param: Let G_1 and G_2 be groups of prime order p , let g be a generator of G_1 and e be an admissible bilinear map from $G_1 \times G_1$ into G_2 . Let two families of functions $\mathcal{H}_1 = \{H_{\bar{k}_1} : \{0, 1\}^{\bar{w}_1} \rightarrow \{0, 1\}^{\bar{n}_1}\}_{\bar{k}_1 \in K_1}$ and $\mathcal{H}_2 = \{H_{\bar{k}_2} : \{0, 1\}^{\bar{w}_2} \rightarrow \{0, 1\}^{\bar{n}_2}\}_{\bar{k}_2 \in K_2}$ be target collision resistant hash functions, where $\bar{w}_1, \bar{w}_2, \bar{n}_1$ and \bar{n}_2 are integers such that $\bar{n}_1, \bar{n}_2 < \log_2 p$ and K_1 and K_2 are key spaces.

Keygen: Choose randoms $x, y, z \in Z_p^*$ and compute $g_1 = g^x, g_2 = g^y, h_1 = g_2^z$ and $Z = e(g, g)$. Select $H_{\bar{k}_1} \in \mathcal{H}_1$ and $H_{\bar{k}_2} \in \mathcal{H}_2$ for some $\bar{k}_1 \in K_1$ and $\bar{k}_2 \in K_2$ respectively where k_1 and k_2 depend on G_1, G_2 and p . For simplicity, denote $H_{\bar{k}_1}$ and $H_{\bar{k}_2}$ as H_1 and H_2 respectively. Then the public key is $\text{PK} = (g, g_1, g_2, h_1, Z, H_1, H_2)$ and the private key is $\text{SK} = (x, y, z)$.

Encryption: To encrypt a message m , first choose randoms $s, t \in Z_p^*$ and compute the following sequentially:

$$c_1 = Z^s m, \quad c_2 = g_2^s, \quad v = H_1(c_1, c_2), \quad \alpha = H_2(g_2^v h_1^t) \quad \text{and} \quad c_3 = (g^\alpha g_1)^s.$$

Then, the ciphertext is $C = (c_1, c_2, c_3, t)$.

Decryption: Upon receipt of ciphertext $C = (c_1, c_2, c_3, t)$, the receiver first computes $v = H_1(c_1, c_2)$ and $\alpha = H_2(g_2^v h_1^t)$. Then, the receiver checks $c_3 = c_2^{y^{-1}(\alpha+x)}$ (or $e(c_3, g_2) = e(g^\alpha g_1, c_2)$). If they are not equal, then output \perp , otherwise compute the plaintext as either

(a) Choose a random $r \in Z_p^*$ and compute

$$m = \frac{c_1}{e(c_3 c_2^r, g^{1/(x+\alpha+yr)})} \quad \text{or}$$

(b) The plaintext is computed as $m = \frac{c_1}{e(c_2^{y^{-1}}, g)}$.

From the above two methods, method (b) is more efficient than method (a) as it performs less exponentiation than method (a). It is noted that if v and α are both replaced by identity and ignore the computation of α , then it is a Boneh-Boyen IBE scheme. Hence, the proposed scheme only increases one multi-exponentiation in both encryption and decryption respectively.

4 Performance Comparisons

In this section, we compare the proposed PKE scheme with a PKE scheme using Canetti-Halevi-Katz's construction (CHK construction) [8] with a strong one-time signature scheme. We choose Boneh-Boyen IBE encryption scheme (BB-E) [2] and Boneh-Boyen strong signature scheme (BB-S) [3] for the CHK construction. The reason is that Boneh-Boyen IBE encryption scheme [2] is same as the proposed PKE scheme and Boneh-Boyen signature scheme is a well-known efficiently secure strong signature scheme. We call this scheme as BB-E/BB-S scheme and describe as follows:

BB-E/BB-S Scheme

Keygen: The key generation is similar to that of section 3 without h_1 and H_2 . Let H_3 be a collision resistant hash function¹. Then, the public key is $PK = (g, g_1, g_2, Z, H_1, H_3)$ and the private key is $SK = (x, y)$.

Encryption: To encrypt a message m , one chooses random integers $a, b, r, s \in Z_p^*$ and computes the following:

$$\begin{aligned} e_0 &= g^a, \quad e_1 = g^b, \quad \alpha = H_1(e_0, e_1), \\ c_1 &= Z^s m, \quad c_2 = g_2^s, \quad c_3 = (g^\alpha g_1)^s, \\ c &= H_3(c_1, c_2, c_3) \quad \text{and} \quad \sigma = g^{1/(a+c+br)}. \end{aligned}$$

Then, the ciphertext is $C = (e_0, e_1, c_1, c_2, c_3, \sigma, r)$.

¹ This is slightly different from a target collision resistant hash function. It is defined (informally) that it is hard to find x and y such that $H_3(x) = H_3(y)$.

Decryption: Upon receipt of the ciphertext $C = (e_0, e_1, c_1, c_2, c_3, \sigma, r)$, the receiver first computes $\alpha = H_1(e_0, e_1)$ and $c = H_1(c_1, c_2, c_3)$, checks $e(\sigma, e_0 \cdot g^c \cdot e_1^r) = Z$ and $e(c_3, g_2) = e(g^\alpha \cdot g_1, c_2)$ (or $c_3 = c_2^{y^{-1}(x+\alpha)}$). If one of the equations are not equal, then output reject symbol \perp , otherwise choose a random $k \in Z_p^*$ and compute the plaintext as either

$$m = \frac{c_1}{e(c_3 c_2^k, g^{1/(x+\alpha+yk)})} \quad \text{or} \quad m = \frac{c_1}{e(c_2^{y^{-1}}, g)}.$$

Let l_1 and l_2 be the length of the representation of an element in G_1 and G_2 respectively. Let $l_p = \log_2 p$, denote exp and m-exp be exponentiation and multi-exponentiation respectively. Then, the performance comparisons are listed in Table 1.

In general, the computation of multi-exponentiation is 1.5 times of an exponentiation. From Table 1, BB-E/BB-S scheme takes 1.5 exponentiation and 1 pairing more than the proposed scheme in encryption and decryption respectively. Hence, the proposed scheme is more efficient than BB-E/BB-S scheme. In order to have a actual timing of the two schemes, we run the simulation using MIRACL software [13] on 3.0GHz Pentium IV computer. The supersingular elliptic curve and size of finite fields (which is 512 bits and p is 160 bits) are chosen to be same as that of [5]. The timing of the two schemes is listed in Table 2.

Table 1. Performance Comparisons

Costs	Public Key Size	Private Key Size	Ciphertext Size	Encryption	Decryption
BB-E/BB-S [4]	$3l_1 + l_2$	$2l_p$	$5l_1 + l_2 + l_p$	5 exp, 1 m-exp	2 exp, 1 m-exp, 2 pairing
Proposed Scheme	$4l_1 + l_2$	$3l_p$	$2l_1 + l_2 + l_p$	2 exp, 2 m-exp	2 exp, 1 m-exp, 1 pairing

Table 2. Timing of PKE Schemes

	Public Key Size	Private Key Size	Ciphertext Size	Encryption	Decryption
BB-E/BB-S [4]	4096 bits	320 bits	6304 bits	15.8ms	55.0ms
Proposed Scheme	5120 bits	480 bits	3232 bits	14.8ms	33.0ms

In Table 2 above, the proposed encryption scheme is more efficient than BB-E/BB-S scheme in term of encryption, decryption and ciphertext size. But the public key size and private key size is slightly longer than that of BB-E/BB-S scheme.

5 Security Analysis

In this section, we give a security proof of the proposed encryption scheme PE1 secure against adaptive chosen ciphertext attacks. Security against adaptive chosen ciphertext attack is defined in the following game.

Definition 5. (Adaptive Chosen Ciphertext Attack (CCA2)) Let $\text{PE} = (\mathcal{P}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ be a public key encryption. Let \mathcal{A} be an attacker modeled as a probabilistic Turing machine. Consider the following game played by a challenger \mathcal{C} and an adversary \mathcal{A} .

Set Up. \mathcal{C} takes a security parameter and runs the key generation algorithm to obtain a public key pk and private key sk . It gives pk to \mathcal{A} and keeps sk secret.

Phase 1. In this phase, \mathcal{A} adaptively makes a number of decryption queries on a ciphertext C . The challenger \mathcal{C} responds with $\mathcal{D}(sk, C)$ or reject.

Challenge. \mathcal{A} outputs two equal length plaintexts (m_0, m_1) . The challenger \mathcal{C} picks a random $b \in \{0, 1\}$, computes a target ciphertext $C^* = \mathcal{E}(pk, m_b)$ and gives it to \mathcal{A} .

Phase 2. The adversary \mathcal{A} continues to make decryption queries on a ciphertext C as in Phase 1 except $C \neq C^*$. The challenger \mathcal{C} responds with $\mathcal{D}(sk, C)$ or reject.

Guess. \mathcal{A} outputs a bit $b' \in \{0, 1\}$. It wins if $b' = b$.

The advantage of an adversary \mathcal{A} is defined as $\text{Adv}_{\text{PE}}^{\text{IND-CCA2}}(\mathcal{A}) = |\Pr[b' = b] - 1/2|$. A secure encryption scheme against adaptive chosen ciphertext attack is defined as follows:

Definition 6. A public key encryption scheme $\text{PE} = (\mathcal{P}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ is said to be (t, q_d, ϵ) -IND-CCA2 secure if the advantage of any t -polynomial time adversary \mathcal{A} is

$$\text{Adv}_{\text{PE}}^{\text{IND-CCA2}}(t, q_d) = \max_{\mathcal{A}} \{ \text{Adv}_{\text{PE}}^{\text{IND-CCA2}}(\mathcal{A}) \} < \epsilon,$$

where the maximum is over all \mathcal{A} which runs in time t and makes at most q_d queries to the decryption oracle.

Before we state the main theorem, we first list the following useful lemma which was defined by Cramer and Shoup in [15].

Lemma 1. ([15], Difference Lemma) Let E_1 , E_2 and F be events defined on some probability space. Suppose that the event $E_1 \wedge \neg F$ occurs if and only if $E_2 \wedge \neg F$ occurs. Then

$$|\Pr[E_1] - \Pr[E_2]| \leq \Pr[F].$$

Theorem 1. The proposed encryption scheme PE1 is (t, q_d, ϵ) -IND-CCA2 secure, assuming that the (t', ϵ') -DBDHI assumption, the (t_1, ϵ_1) -target collision resistant hash function H_1 and the (t_2, ϵ_2) -target collision resistant hash function H_2 hold, such that

$$\epsilon \leq \epsilon' + \epsilon_1 + 2\epsilon_2 + \frac{qd}{p},$$

where t' , t_1 and t_2 are essentially the same as t and $q_d < p$.

Proof. This theorem is proved based on reductionist proof, that is, suppose an adversary \mathcal{A} could break the proposed encryption scheme PE1 in time t with advantage ϵ , the goal is to construct an algorithm \mathcal{B} which solves the DBDHI problem in time t' with advantage ϵ' . First, the algorithm \mathcal{B} is given an instance $(g, A_1, \dots, A_q, T) \in G_1^{q+1} \times G_2$, where $A_i = g^{\mu^i} \in G_2$ for $i = 1, \dots, q$ and some unknown $\mu \in Z_p^*$. The algorithm \mathcal{B} 's goal is to output 1 if $T = e(g, g)^{1/\mu}$ and 0 otherwise. Now, we construct the algorithm \mathcal{B} as a simulator which interacts with \mathcal{A} in the IND-CCA2 game as follows:

- Preparation: First, the algorithm \mathcal{B} constructs $\tilde{g} \in G_1$ and $Z, T_1 \in G_2$ as follows:
 P1. Choose randoms $w_1, \dots, w_{q-1} \in Z_p^*$ and let $f(z)$ as $f(z) = \prod_{i=1}^{q-1} (z + w_i)$, then $f(z)$ can be written as $f(z) = \sum_{i=0}^{q-1} c_i z^i$, where the constant c_0 is non-zero.
 P2. Compute $\tilde{g} = \prod_{i=0}^{q-1} A_i^{c_i} = g^{f(\mu)}$, $\hat{g} = \prod_{i=0}^{q-1} A_{i+1}^{c_i} = g^{\mu f(\mu)} = \tilde{g}^\mu$ and $Z = e(\tilde{g}, \tilde{g})$.
 P3. If $\tilde{g} = 1$, then there exists some j such that $w_j = -\mu$. In this case, the algorithm \mathcal{B} would be able to solve BDHI problem and hence DBDHI problem. Hence, we assume that $w_i \neq -\mu$ for $1 \leq i \leq q-1$.
 P4. \mathcal{B} computes $T_1 = T^{c_0^2} \cdot T_0$, where

$$T_0 = \prod_{i=0}^{q-2} \prod_{j=1}^{q-1} e(A_i, A_j)^{c_{i+1}c_j} = \prod_{i=0}^{q-2} \prod_{j=1}^{q-1} e(g^{\mu^i}, g^{\mu^j})^{c_{i+1}c_j} \text{ with } A_0 = g.$$

It is easily checked that $T_1 = e(g^{f(\mu)/\mu}, g^{f(\mu)/\mu}) = e(\tilde{g}, \tilde{g})^{1/\mu}$ if $T = e(g, g)^{1/\mu}$, otherwise T_1 is random in $G_2 \setminus \{T_0\}$.

P5. Let H_1 and H_2 be sub-function of \mathcal{H}_1 and \mathcal{H}_2 with key index K_1 and K_2 by the two cyclic groups G_1, G_2 and p respectively.

Key Set Up: \mathcal{B} chooses randoms $k, a, z \in Z_p^*$ and computes the following sequentially

$$\alpha^* = H_2(\hat{g}^k), \quad b = \alpha^* a^{-1} \quad \text{and} \quad \tilde{h}_1 = \tilde{g}_2^z.$$

As the algorithm \mathcal{B} does not know μ , we implicitly define $x = -a(\mu + b)$ and $y = \mu$ so that $\tilde{g}_1 = \hat{g}^{-a} \tilde{g}^{-\alpha^*} = \tilde{g}^{-a(\mu+b)} = \tilde{g}^x$ and $\tilde{g}_2 = \hat{g} = \tilde{g}^\mu = \tilde{g}^y$. Then the algorithm \mathcal{B} gives the public key PK = $(\tilde{g}, \tilde{g}_1, \tilde{g}_2, \tilde{h}_1, Z, H_1, H_2)$ to \mathcal{A} and keeps the private key SK = (x, y, z) secret, where x and y are unknown to \mathcal{B} .

Phase 1: The adversary \mathcal{A} makes a number of decryption queries. If the adversary submits a ciphertext $C = (c_1, c_2, c_3, t)$ for decryption, \mathcal{B} first computes $v = H_1(c_1, c_2)$ and $\alpha = H_2(\tilde{g}_2^v \tilde{h}_1^t)$; and checks $\alpha = \alpha^*$. If there are equal, then the simulation aborts (as \mathcal{B} is not able to decrypt the ciphertext), otherwise \mathcal{B}

checks $e(\tilde{g}^\alpha \tilde{g}_1, c_2) = e(c_3, \tilde{g}_2)$. If they are not equal, then return \perp , otherwise generates a decryption key $DK = (d_1, d_2)$ as follows:

Ph1. Randomly pick w_i from $\{w_1, \dots, w_{q-1}\}$ for $1 \leq i \leq q-1$ such that w_i is not chosen before.

Ph2. \mathcal{B} computes $r \in Z_p$ such that $(r-a)(\mu+w_i) = \alpha+x+ry$. Substitute x and y into the equation, we have

$$(r-a)(\mu+w_i) = \alpha - a(\mu+b) + r\mu$$

and obtain $r = a + \frac{\alpha-ab}{w_i} \in Z_p$.

Ph3. Then, the decryption key is $d_2 = r$ and

$$d_1 = (\tilde{g}^{1/(\mu+w_i)})^{1/(r-a)} = \tilde{g}^{1/(\mu+w_i)(r-a)} = \tilde{g}^{1/(\alpha+x+ry)}.$$

Finally, \mathcal{B} returns the plaintext m as follows:

$$m = \frac{c_1}{e(c_3 c_2^{d_2}, d_1)}.$$

Challenge: After the number of queries in **Phase 1**, \mathcal{A} outputs two equal length messages m_0 and m_1 on which it wishes to be challenged. \mathcal{B} flips a fair coin $b \in \{0, 1\}$, chooses a random number $\gamma \in Z_p^*$ and computes

$$c_1^* = T_1^\gamma \cdot m_b, c_2^* = \tilde{g}^\gamma, c_3^* = \tilde{g}^{-a\gamma}, v^* = H_1(c_1^*, c_2^*), t^* = (k - v^*)z^{-1} \bmod p.$$

Then, \mathcal{B} responds the challenge ciphertext $C^* = (c_1^*, c_2^*, c_3^*, t^*)$. It can be checked that C^* is valid encryption of m_b by first defining $s^* = \gamma/\mu$. If $T_1 = e(\tilde{g}, \tilde{g})^{1/\mu}$, then

$$\begin{aligned} T_1^\gamma &= e(\tilde{g}, \tilde{g})^{\gamma/\mu} = e(\tilde{g}, \tilde{g})^{s^*}, \\ c_2^* &= \tilde{g}^\gamma = \tilde{g}_2^{\gamma/\mu} = \tilde{g}_2^{s^*}, \\ c_3^* &= \tilde{g}^{-a\gamma} = \tilde{g}^{-a\mu(\gamma/\mu)} = \tilde{g}^{(x+\alpha^*)(\gamma/\mu)} = (\tilde{g}^{\alpha^*} \cdot \tilde{g}_1)^{s^*}. \end{aligned}$$

On the other hand, if T_1 is random in $G_2 \setminus \{T_0\}$, then C^* is independent of the bit b in the adversary's view.

Phase 2: The adversary \mathcal{A} continues to make the decryption queries on ciphertext C similar to **Phase 1** except $C = C^*$.

Guess: After the number of decryption queries, the adversary \mathcal{A} returns a bit $b' \in \{0, 1\}$. If $b \neq b'$, the algorithm \mathcal{B} return $\beta' = 0$, else it returns $\beta' = 1$. This completes the description of the simulator. Note that the simulator behaves exactly the same as in the original public key encryption except the abortion in **Phase 1**, we will discuss this below.

Analysis: We analyse the success probability of \mathcal{B} by considering a sequence of the "indistinguishable" modified games from game G_0 to game G_3 , where G_0 is the original game and the last game G_3 gives no advantage to the adversary \mathcal{A} . Let $b' \in \{0, 1\}$ be the output of \mathcal{A} and E_i be the event that $b' = b$ in the game G_i for $0 \leq i \leq 3$. Then, we have

$$\text{Adv}_{\text{PE1}}^{\text{IND-CCA2}}(\mathcal{A}) = |\Pr[E_0] - 1/2|$$

and the sequence of games are described as follows:

Game G_1 : First, game G_0 is modified to a new game G_1 such that the decryption oracle in **Phase 1** is modified with the rejection rule as follows : If the adversary submits a ciphertext $C = (c_1, c_2, c_3, t)$ with $\alpha = \alpha^*$, where $\alpha = H_1(\tilde{g}_2^c \tilde{h}_1^t)$ and $v = H_1(c_1, c_2)$, the decryption oracle immediately outputs reject and halt. Since the adversary has no information (in a statistical sense) about α^* from the challenge ciphertext C^* , hence the chance of having $\alpha = \alpha^*$ is ϵ_2 . Therefore, by Lemma 1, we have $|\Pr[E_1] - \Pr[E_0]| \leq \epsilon_2$.

Game G_2 : To turn game G_1 to a new game G_2 , the decryption oracle in **Phase 2** is modified such that the rejection rule is applied as follows: If the adversary \mathcal{A} submits a ciphertext $C = (c_1, c_2, c_3, t)$ with $(c_1, c_2) \neq (c_1^*, c_2^*)$ such that either $(v = v^*, t = t^*)$ or $\alpha = \alpha^*$ where $v = H_1(c_1, c_2)$ and $\alpha = H_2(\tilde{g}_2^v \tilde{h}_1^t)$, then the decryption oracle immediately outputs reject and halt. The chance of having the above two cases are ϵ_1 and $\epsilon_2 + \frac{qd}{p}$ respectively². Hence, by Lemma 1, we have $|\Pr[E_2] - \Pr[E_1]| \leq \epsilon_1 + \epsilon_2 + \frac{qd}{p}$.

Game G_3 : In this game, the encryption oracle is modified so that c_1^* is replaced by random c'_1 in G_2 . Due to this change, c'_1 does not depend on T , then Game G_3 and Game G_2 are equal unless the adversary \mathcal{A} can distinguish $e(g, g)^{1/\mu}$ from the random element in G_2 . Hence, we have $|\Pr[E_3] - \Pr[E_2]| \leq \epsilon'$. Furthermore, since c'_1 is independent of the challenge bit b and does not provide any information in the adversary's view, therefore we have $\Pr[E_3] = 1/2$.

Combine the results from the above games, we immediately obtain the following:

$$\epsilon \leq \epsilon' + \epsilon_1 + 2\epsilon_2 + \frac{qd}{p}.$$

6 Conclusion

In this paper, we constructed an efficient public key encryption with publicly verifiable. The proposed public key encryption scheme is secure against adaptive chosen ciphertext attack based on the hardness of decisional bilinear Diffie-Hellman inversion assumption and the target collision resistance hash functions. Furthermore, the computational complexity is almost as efficient as Boneh-Boyen IBE encryption scheme with only an increase of one multi-exponentiation in both encryption and decryption.

References

- [1] D. Boneh and X. Boyen, "Secure identity based encryption without random oracles," *Advances in Cryptology - Crypto'04, Lecture Notes in Computer Science vol.3152*, pp.443-459, Springer-Verlag, 2004.
- [2] D. Boneh and X. Boyen, "Efficient selective-id secure identity based encryption without random oracles," *Advances in Cryptology - Eurocrypt'04, Lecture Notes in Computer Science vol.3027*, pp. 223-238, Springer-Verlag, 2004.

² Due to the page limit, the proof of these are given in the full paper.

- [3] D. Boneh and X. Boyen, "Short signatures without random oracles," *Advances in Cryptology - Eurocrypt'04*, Lecture Notes in Computer Science, vol.3027, pp.56-73, Springer-Verlag, 2004.
- [4] D. Boneh, R. Canetti, Shai Halevi, and J. Katz, "Chosen-Ciphertext Security From Identity-Based Encryption," Accepted to *SIAM Journal on Computing*. Available from <http://www.cs.umd.edu/~jkatz/papers/id-cca-journal/pdf>.
- [5] D. Boneh and M. Franklin, "Identity-based encryption from Weil pairing," *SIAM J. Comput.*, vol.32, no.3, pp.586-615, 2003.
- [6] D. Boneh and J. Katz, "Improved efficiency for CCA-secure cryptosystems built using identity based encryption," *Topics in Cryptology – CT-RSA 2005*, Lecture Notes in Computer Science vol.3376, pp. 87-103, Springer-Verlag, 2005.
- [7] X. Boyen, Q. Mei, and B. Waters, "Direct chosen ciphertext security from identity-based techniques," In *ACM Conference on Computer and Communications Security CCS 2005*, pp. 320-329, ACM Press, 2005. Full version available at <http://eprint.iacr.org/2005/288>.
- [8] R. Canetti, S. Halevi and J. Katz, "Chosen-ciphertext security from identity-based encryption," *Advances in Cryptology - Eurocrypt'04*, Lecture Notes in Computer Science vol.3027, pp. 207-222, Springer-Verlag, 2004.
- [9] R. Cramer and V. Shoup, "A practical public key cryptosystem provably secure against adaptive chosen ciphertext attack," *Advances in Cryptology - Crypto'98*, Lecture Notes in Computer Science vol.1462, pp. 13-25, Springer-Verlag, 1998.
- [10] R. Cramer and V. Shoup, "Design and analysis of practical public-key encryption schemes secure adaptive chosen ciphertext attack," *SIAM J. Comput.*, vol.33, no.1, pp.167-226, 2003.
- [11] D. Dolev, C. Dwork, and M. Naor, "Non-malleable cryptography," *The 23rd Annual ACM Symposium on Theory of Computing – STOC'91*, pp.542-552, ACM press, 1991.
- [12] E. Kiltz, "On the limitation of the spread of an IBE-to-PKE transformation," *Public key Cryptography - PKC'06*, Lecture Notes in Computer Science vol.3958, pp. 274-289, Springer-Verlag, 2006.
- [13] MIRACL, Multiprecision integer and rational arithmetic C/C++ library, Shamus Software Ltd. Available from <http://indigo.ie/~mscott/>.
- [14] C. Rackoff and D. Simon, "Non-interactive zero-knowledge proof of knowledge and chosen ciphertext attack," *Advances in Cryptology - Crypto'91*, Lecture Notes in Computer Science Vol.576, pp.46-64, Springer-Verlag, 1991.
- [15] V. Shoup, "Sequences of games: a tool for taming complexity in security proofs," manuscript, 2004. Available from <http://eprint.iacr.org/2004/332>.
- [16] C. H. Tan, "Chosen ciphertext security from identity-based encryption without strong condition," *Workshop on Security*, Lecture Notes in Computer Science vol. 4266, pp.296-311, Springer-Verlag, 2006.
- [17] B. Waters, "Efficient identity-based encryption without random oracles," *Advances in Cryptology - Eurocrypt'05*, Lecture Notes in Computer Science vol.3494, pp.114-127, Springer-Verlag, 2005.

Solving Bao's Colluding Attack in Wang's Fair Payment Protocol

M. Magdalena Payeras-Capellà, Josep L. Ferrer Gomila, and Llorenç Huguet Rotger

Universitat de les Illes Balears.
Carretera de Valldemossa, km7,5. 07120 Palma de Mallorca, Spain
{mpayeras, jlferrer, l.huguet}@uib.es

Abstract. An electronic purchase is an essential operation of electronic commerce. Fairness in the exchange of money and product, as well as anonymity of the buyer, are desirable features. In Asiacrypt 2003, C.H. Wang [8] presented a purchase protocol satisfying both anonymity and fairness, adapting the anonymous payment system of Brands [2], using a restrictive confirmation signature scheme. Later, In Asiacrypt 2004, Feng Bao [1] demonstrated that Wang's protocol [8] can be vulnerable to attacks produced by colluding users, and he affirmed that the protocol cannot be corrected due to the anonymity of the protocol. We will show that it is possible to correct Wang's protocol in order to avoid colluding attacks. We present a solution that modifies slightly the original protocol, maintaining the anonymity and untraceability of the original version. Finally, we discuss the convenience to achieve the property of timeliness.

1 Introduction

Some electronic services require an exchange of elements between two or more users. A fair exchange of values always provides an equal treatment to all the users, and, at the end of the execution of the exchange, all parties have the element that wished to obtain, or the execution has not been solved successfully (in this case nobody has the expected element).

Among the electronic applications that require a fair exchange of information we can find electronic contract signing, certified electronic mail and electronic purchase (payment in exchange for a receipt or a digital product).

The exchange of a payment and a receipt (or a digital product) occurs in an electronic purchase, in which different kinds of payment systems can be used. Even though the payment is executed during the purchase operation, the buyer doesn't obtain the tangible product until it is delivered to him. The confidence in the service is a fundamental aspect, and the user will be more motivated to make the payment if he receives a receipt that will demonstrate (without possible repudiation on the part of the seller) that the user has made the payment. In the purchase of a digital product, the exchange is slightly different. Now it is not necessary that the seller sends a receipt to the buyer, instead he will send the digital product directly, but the fairness of the exchange is still required since the buyer doesn't want to pay if he isn't sure that he will receive the product, whereas the salesman doesn't want to send the product before receiving the payment.

The fair exchange of a coin for a receipt or for a product (also called fair payment or fair purchase) is required in the use of electronic cash systems [2, 5, 6, 8] for the purchase of goods or services offered electronically. In payments using credit cards, instead, the purchase order including the card number (and the signature) is exchanged for the receipt or digital good. Although the fair payment using credit card can be considered an application of the contract signing protocols, fair payments using electronic cash, where the coins are part of the exchange (element provided by the buyer), require specific exchange protocols, and cannot be considered an application of the contract signing protocols due to some specific features. The interruption of the exchange can result in the loss of a coin for both parties or the loss of anonymity of a pretended anonymous user. As an example, in an off-line anonymous system, if the buyer does not know if the seller has received the coin, he cannot spend the coin again, because if the seller has received the coin the buyer would be identified and accused of double spending of coins.

The interruptions can be due to net failures or fraudulent behavior. Consequently, it is possible that the buyer provides the coin and do not obtain the good or receipt from the seller, or that the seller sends the good and do not receive the coin. Atomicity [7] allows linking a group of operations so they must be executed totally or not executed at all.

Anonymity is a desired feature in payments [2, 3, 5, 8]. For this reason it is interesting that the purchase protocols maintain the anonymity of the payment protocol, so adding atomicity [3, 4] to an anonymous payment system we will have an anonymous fair payment protocol.

A trusted third party (TTP) can be used to solve conflicts between the users if the exchange is not completed, but it is desirable that the TTP doesn't participate in each protocol run [5]. In a fair payment protocol, it is also desirable that the bank doesn't check the coin during the payment (off-line payment).

Brands payment system [2] is both anonymous and off-line. Adding atomicity to Brands protocol will result in an anonymous fair exchange protocol. In [8], Wang presented a fair payment protocol that can be used with Brands payment system. In section 2, Wang's protocol is described, including Brands payment protocol and the application of Wang's fair payment protocol to Brands system. Section 3 enunciates Bao's attack [1] to Wang's protocol. Section 4 presents the modification to Wang's protocol to solve the vulnerability. Finally, section 5 includes an analysis of the modified protocol discussing the convenience of an asynchronous approach.

2 Wang's Protocol

Wang's protocol [8] involves four parties: the buyer (U), the seller (M), the bank (B) and the trusted third party (TTP). According to Wang, in an e-cash system fairness cannot be achieved because the buyer must send true electronic coins (*e-coins*) to the seller. For this reason, he decides to use *pseudo e-coins* that can be converted to true *e-coins* by the TTP. The buyer applies a *Restrictive Confirmation Signature Scheme* (RCSS) to sign the purchase agreement that includes the names of both buyer and seller, price of the item, date of the purchase and other parameters. The RCSS provides anonymity to the system protecting the buyer purchase information, because the RCSS restricts the capacity of the parties to confirm it.

$Sign_{DCS}(S, C, m)$, represents a signature of S over the message m that can be confirmed by C . $G = \{V_i\}_{i=1,\dots,n}$ are a group of verifiers predetermined by S . Then $Sign_{RCSS}(S, C, G, m)$ is a *Restrictive Confirmation Signature* over m if C can convince only a specific group of verifiers $V_i \in G$ of the validity of $Sign_{RCSS}(S, C, G, m)$.

The protocol is formed by three procedures: *withdrawal*, *payment* and *deposit*. If the TTP is required, an additional procedure (*dispute resolution*) can be executed.

- **Withdrawal.** The buyer U withdraws the money from the bank B and obtains an *e-coin*. A blind signature is applied to obtain anonymity.
- **Payment.** The buyer U and the seller M exchange an electronic coin and a product (or receipt). U and M have a purchase agreement, a document that contains the price and the description of the product. The buyer sends enough *pseudo e-coins* and the signature $RCSS$ over the agreement to M . The buyer doesn't send true *e-coins* until the product has been checked.
 1. U selects the product from M 's web and signs a purchase agreement:

$$\theta = Sign_{RCSS}(U, M, TTP, OA)$$

$$OA = \{ID_U, ID_M, \text{data/purchase information, product description, coin parameters}\}$$
 2. U sends the *pseudo e-coins* and θ to M .
 3. M checks the *pseudo e-coins* and θ . If both are valid, M sends the product to U . If the purchase doesn't conclude successfully, M can prove the validity of θ to the TTP (thanks to the $RCSS$) and requests the conversion of the *pseudo e-coins* to true *e-coins*.
 4. U checks the product received from M . If the product is correct, U sends the true *e-coins* to M .
- **Disputes.** Two kinds of disputes can arise between the parties. In the first case, M can refuse to send the product to U or try to cheat with a false product. Then, U will not send the true *e-coins* to M . In the second case, U can refuse to send the true *e-coins* to M after the reception of the product. Then, M will contact with the TTP executing the dispute resolution procedure to convert the *pseudo e-coins* into true *e-coins*.
 1. M sends the *pseudo e-coins*, the order agreement OA and the signed order agreement θ to the TTP and proves that θ is a valid signature of U . Nobody except M and TTP can be convinced of the validity of θ , because a $RCSS$ is used in the construction of θ ($\theta = Sign_{RCSS}(U, M, TTP, OA)$).
 2. M sends the product to the TTP. The TTP checks the description of the product in OA . Then, the TTP sends to M a transformation certificate (T_{Cer}) that can be used for the conversion of the *pseudo e-coins* into true *e-coins*.
- **Deposit.** Normally, M sends true *e-coins* for deposit. Alternatively, he can deposit *pseudo e-coins* together with their transformation certificate (T_{Cer}).

2.1 Brand's Payment Protocol

This section includes a summary of Brands payment system [2], that later will be used to show how Wang's fair payment protocol works.

- **Initialization.** Let p and q be two large primes. The *Bank* publishes a generator tuple (g, g_1, g_2) in G_q ($G_q \square Z_p^*$) and two collision resistant hash functions: $H: G_q \times G_q \times G_q \times G_q \times G_q \rightarrow Z_q^*$ and $H_0: G_q \times G_q \times ID \times DATE/TIME \rightarrow Z_q^*$. The *Bank*

generates a random number $x_B \in Z_q^*$ as his secret key corresponding to the public key $y_B (y_B = g^{x_B} \text{ mod } p)$.

- **Account opening.** The buyer U randomly selects $u_1 \in Z_q^*$ and sends $I = g_1^{u_1} \text{ mod } p$ to the *Bank* if $I g_2 \neq 1$. The identifier I is used as an account number for U . The *Bank* publishes the values $g_1^{x_B} \text{ mod } p$ and $g_2^{x_B} \text{ mod } p$ so that U can compute $z = (I g_2)^{x_B} \text{ mod } p$.
- **Withdrawal.** U withdraws an amount of money from his account.
 1. U shows to the *Bank* that he is the owner of his account signing a request.
 2. The *Bank* generates a random w , and sends $e_1 = g^w \text{ mod } p$ and $e_2 = (I g_2)^w \text{ mod } p$ to U .
 3. U selects randomly $x_1, x_2, u, v, s \in Z_q^*$ and calculates $A = (I g_2)^s \text{ mod } p$, $B = g_1^{x_1} g_2^{x_2} \text{ mod } p$, $z' = z^s \text{ mod } p$, $e_1' = e_1^u g^v \text{ mod } p$, $e_2' = e_2^{su} A^v \text{ mod } p$, $c' = H(A, B, z', e_1', e_2') \text{ mod } q$ and sends $c = c' / u \text{ mod } q$ to the *Bank*.
 4. The *Bank* sends $r = cx + w \text{ mod } q$ to U .
 5. U calculates $r' = ru + v \text{ mod } q$ and verifies $g^r = y_B^c e_1 \text{ mod } p$ and $(I g_2)^r = z^c e_2 \text{ mod } p$.
- **Payment.** U pays the seller (M) an amount of money.
 1. U sends to M : A, B , and its signature $\text{Sign}(A, B) = (g^{r'} = y_B^{H(A, B, z', e_1', e_2')}, e_1', A^{r'} = z'^{H(A, B, z', e_1', e_2')}, e_2', (z', e_1', e_2', r'))$.
 2. If $A \neq 1$, M sends $d = H_0(A, B, I_M, \text{date/time})$ to U .
 3. U sends $r_1 = d(u_1 s) + x_1 \text{ mod } q$ and $r_2 = ds + x_2 \text{ mod } q$ to M .
 4. M verifies the signature of the *Bank* over the coin: $\text{Sign}(A, B)$, $g^{r'} = e_1' * y_B^c \text{ mod } p$, $A^{r'} = e_2' * z'^c \text{ mod } p$, and accepts the payment if $g_1^{r_1} g_2^{r_2} = A^d B$.
- **Deposit.** M deposits an electronic coin in his account.
 1. M sends the coin to the *Bank*: $A, B, \text{Sign}(A, B), (r_1, r_2)$ and (date/time) .
 2. If $A = 1$, then the *Bank* doesn't accept the transaction.
 3. If it is not the case, the *Bank* calculates d and verifies the signature $\text{Sign}(A, B)$ and $g_1^{r_1} g_2^{r_2} = A^d B$. The *Bank* searches A in the database. There are two cases:
 - A doesn't appear in the database; then the *Bank* stores the information of the transaction and increases the amount in M 's account.
 - A appears in the database; this is a fraud attempt. If the stored information shows that the deposit was done by M and date/time is the same that the value in the new transaction, then M is depositing the coin again.
 If the values are different, then the coin have been double spent, if (d, r_1, r_2) are the information of the new transaction and (d', r_1', r_2') are the information of the stored information, the *Bank* can calculate $g_1^{(r_1 - r_1') / (r_2 - r_2')}$, the account number of the double spender.

2.2 Wang's Protocol Applied to Brands Payment System

In this section the technique of the *pseudo e-coin* described by Wang is applied to Brands protocol, modifying the original protocol to achieve atomicity.

• Withdrawal

1. U randomly selects t_c in Z_q^* and calculates $(a_c, b_c) = (g^{t_c} \text{ mod } p, y_{TTP}^{t_c} \text{ mod } p)$, a pair of confirmation parameters. y_{TTP} and x_{TTP} represent the public and private key of the TTP, respectively. The element c' is modified, now $c' = H(A, B, z', e_1', e_2', b_c) + a_c \text{ mod } q$.
2. The element $\langle A, B, (z', e_1', e_2', r', a_c, b_c) \rangle$ represents a *pseudo e-coin*.

- **Payment**

1. U signs the purchase agreement, $\theta = \text{Sign}_{\text{RCSS}}(U, M, \text{TTP}, OA)$, where $OA = \{ID_U, ID_M, \text{data} / \text{purchase information, product description, } (A, B)\}$
2. U sends a new *pseudo e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c) \rangle$ to M .
3. M verifies the *pseudo e-coin* and θ . If they are valid and $A \neq 1$, then he sends the challenge $d = H_0(A, B, ID_M, \text{date/time})$ to U .
4. U sends the responses $r_1 = d(u_1s) + x_1 \bmod q$ and $r_2 = ds + x_2 \bmod q$ to M .
5. M will accept the *pseudo e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$, if the following can be verified:

$$\begin{aligned} g^{r'} &= y_B^{H(A, B, z', e_1, e_2, bc) + ac} e_1, \\ A^{r'} &= z^{H(A, B, z', e_1, e_2, bc) + ac} e_2, \\ g_1^{r_1} g_2^{r_2} &= A^d B \end{aligned}$$

If it is the case, M sends the product to U .

6. U checks the product sent by M . If it is correct, U sends t_c to M . The group of elements $\langle A, B, (z', e_1', e_2', r', a_c, b_c, t_c), (d, r_1, r_2) \rangle$ represents a true *e-coin* that can be deposited because $a_c = g^{t_c} \bmod p$ and $b_c = y_{\text{TTP}}^{t_c} \bmod p$.

- **Disputes.** If U refuses to send t_c to M , M will contact the TTP .

1. M sends the purchase agreement OA , the signature θ , the product and the *pseudo e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$ to the TTP .
2. The TTP checks the product, the *pseudo e-coin* and the signature θ . If they are valid, the TTP sends a transformation certificate $T_{\text{Cer}} = (E_c, T_c)$, to M , where $E_c = a_c^\sigma \bmod p$ (σ is a random selected by the TTP) and $T_c = \sigma + x_{\text{TTP}} F(a_c, E_c) \bmod q$. The transformation certificate can be used to verify the relation between a_c and b_c using $a_c^{T_c} \stackrel{?}{=} E_c b_c^{F(a_c, E_c)} \bmod p$.
3. The TTP sends the product to U .

- **Deposit.** In the normal case, M sends the true *e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c, t_c), (d, r_1, r_2) \rangle$, to the bank. If U aborts the payment process, M can get T_{Cer} , in this case, the *pseudo e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$ together with the transformation certificate $T_{\text{Cer}} = (E_c, T_c)$, are valid elements to be deposited.

3 Bao's Colluding Attack

In [1], Feng Bao presents three attacks done by colluding users. Two of these attacks are against signature exchange protocols and the other is against fair payment protocols. The last attack can be applied to fair payment systems, where the users can contact a TTP to solve an uncompleted exchange. Wang's protocol is vulnerable to this attack, and [1] includes a description of the attack to this protocol.

To attack the system, the seller M colludes with a new user, a conspirator called C . The fair payment begins its execution between U and M . After receiving the *pseudo e-coin* from the buyer U , M sends the *pseudo e-coin* to the TTP but claims that C sent him the money. Then, the TTP converts the *pseudo e-coin* to a true *e-coin* for M and sends the product to C . U will not obtain any element, since the TTP doesn't know his involvement in the exchange.

The resulting payment process is as follows:

1. The malicious seller M executes the payment protocol with U until step 5 obtaining the *pseudo e-coin* without delivering the product.

2. M colludes with C who signs a forged order agreement between M and C . The new agreement is called $\theta' = \text{Sign}_{\text{RCSS}}(C, M, TTP, OA')$, where $OA' = \{ID_C, ID_M, \text{information/date, description of the product, (A,B)}\}$. This false agreement order only differs with the original in the first signed element (now ID_U has been substituted by ID_C).
3. M begins the dispute process sending the new agreement OA' and the RCSS signature θ' over OA' , the product and the *pseudo e-coin* $\langle A, B, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$ to the TTP . The TTP cannot distinguish between the real agreement and the new one because of the anonymity of the payment system. If $d = H(A, B, ID_M, \text{date/time})$ would include ID_U , the attack would not be possible, but the anonymity would disappear.
4. The TTP converts the *pseudo e-coin* into a true *e-coin* for M and sends the product to C . U doesn't obtain anything.

Bao states that the problem of the fair payment is resulting from the use of electronic coins generated using the bank's private key, without any relation with the buyer's identity. Bao states that Wang's protocol cannot be fixed incorporating new elements to indicate that the exchange is between U and M . A purchase pre-contract cannot be used to indicate that the trade is between U and M since the conspirator C can just simulate U by doing everything U does. Due to the anonymity and untraceability of the system, no one can distinguish C from U .

4 Solving Bao's Attack to Wang's Protocol

In this section we will show how Wang's protocol can be corrected to avoid colluding attacks without losing the anonymity and untraceability of the payment system. Two of the features of Wang's protocol can be considered the reasons of the vulnerability. The first one is the relation between the order agreement θ and the *e-coin* (or the *pseudo e-coin*). In the original protocol it is possible to substitute θ and state that it is related with a payment including the *e-coin*.

The second one is a consequence of the synchrony of the protocol and the fact that the buyer never contacts with the TTP . In a timeliness protocol a party can contact the TTP to know the final state of an exchange whenever he wants, so timeliness is not achieved in Wang's protocol. Instead, the buyer waits for a possible message from the TTP . The buyer cannot claim a dispute if he doesn't receive any element from the TTP after a colluding attack. Moreover, if the TTP is cheated, he will contact with the colluding conspirator and will not contact with U .

A possible solution to the vulnerability is the inclusion in the *e-coin* of a new element. This new element acts a link between θ and the *e-coin*. The modified protocol is as follows.

• Modified Withdrawal

1. U randomly selects a new element α (secret element) and t_c in Z_q^* and calculates $(a_c, b_c) = (g^{t_c \text{ mod } p}, y_{TTP}^{t_c \text{ mod } p})$, and a new element $D = g_1^\alpha$. The element c' is now modified to incorporate D , $c' = H(A, B, D, z', e_1', e_2', b_c) + a_c \text{ mod } q$.
2. Now the coin is $\langle A, B, D, (z', e_1', e_2', r', a_c, b_c) \rangle$ and is related with a secret element α . Since c' is used to calculate r' , D cannot be substituted in the coin.

- **Modified Payment**

1. U selects the product, signs a purchase agreement $\theta = \text{Sign}_{\text{RCSS}}(U, M, \text{TTP}, OA)$, where $OA = \{ID_U, ID_M, \text{data/purchase information, product description, } (A, B)\}$, and sends it to M .
2. U sends a new *pseudo e-coin* $\langle A, B, D, (z', e_1', e_2', r', a_c, b_c) \rangle$ to M .
3. U sends $r_1 = d(u_1s) + x_1 \text{ mod } q$ and $r_2 = ds + x_2 \text{ mod } q$, to M . U creates the new elements: $d_2 = H(A, c')$ and a random called d_3 . Then U calculates $P = g_1^{d_3}$, and $r_3 = d_2 * \alpha + \theta * d_3 \text{ mod } q$, and sends d_2, P and r_3 to M .
4. M verifies the *pseudo e-coin* and θ , and checks its relation using $g_1^{r_3} = P^\theta * D^{d_2}$. With this verification M can be sure that U knows the secret element α . This element will be used by the TTP to avoid colluding attacks. If all of them are valid and $A \neq 1$, then sends the challenge $d = H_0(A, B, ID_M, \text{date/time})$ to U .
5. M will accept the *pseudo e-coin* $\langle A, B, D, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$, if the following can be verified:

- $g^{r'} = y_B^{H(A, B, z', e_1, e_2, bc) + ac} e_1'$,
- $A^{r'} = z'^{H(A, B, z', e_1, e_2, bc) + ac} e_2'$,
- $g_1^{r_1} g_2^{r_2} = A^d B$

If it is the case, M sends the product to U .

6. U checks the product sent by M . If it is correct, U sends t_c to M . Because $a_c = g^{t_c} \text{ mod } p$ and $b_c = y_{\text{TTP}}^{t_c} \text{ mod } p$, $\langle A, B, D, (z', e_1', e_2', r', a_c, b_c, t_c), (d, r_1, r_2) \rangle$ represents a true *e-coin* that can be deposited.

- **Modified dispute resolution.** If U refuses to send t_c to M , M will contact the TTP.

1. M sends the purchase agreement OA , the signature θ , the product and the *pseudo e-coins* $\langle A, B, D, (z', e_1', e_2', r', a_c, b_c), (d, r_1, r_2) \rangle$, to the TTP.
2. The TTP checks the product, the *pseudo e-coin*, the signature θ , and now also the relation between the *pseudo e-coin* and θ . The element $r_3 = d_2 \alpha + \theta d_3 \text{ mod } q$ can be used to prove that the new element of the coin ($D = g_1^{\alpha}$) is calculated from the secret element (α) included in it. If all of them are valid, the TTP sends a transformation certificate $T_{\text{Cer}} = (E_c, T_c)$, to M , where $E_c = a_c^\sigma \text{ mod } p$ (σ is a random number selected by the TTP) and $T_c = \sigma + x_{\text{TTP}} F(a_c, E_c) \text{ mod } q$. The transformation certificate can be used to verify the relation between a_c and b_c using $a_c^{T_c} \stackrel{?}{=} E_c b_c \text{ mod } p$.
3. The TTP sends the product to U .

5 Analysis of the Modified Protocol

Section 4 presents a modification of Wang's protocol. Now, we will explain how the modified protocol cannot be flawed by colluding users. A new element D has been added to the coin and another element, r_3 , is used to link the order agreement θ and the e-coin.

D and α are related elements, D can be calculated from α , but α cannot be calculated from D . Similarly, r_3 can be calculated from θ and θ cannot be calculated from r_3 , and for these reasons r_3 cannot be calculated without the knowledge of α .

If C and M are colluding users, and M sends to C all the information available, C can try to generate a false agreement $\theta' = \text{Sign}_{\text{RCSS}}(C, M, TTP, OA')$, where $OA' = \{ID_C, ID_M, \text{information/date, description of the product, (A,B)}\}$. C can generate OA' and therefore θ' since there isn't any secret element in them. However, C doesn't know the value of α , and for this reason C cannot create a false linking element $r_3' = d_2 * \alpha + \theta' * d_3 \text{ mod } q$. The use of an incorrect α would be detected by the TTP, since the TTP checks if the *pseudo e-coin* and θ' are related through $g_1^{r_3} = P^\theta * D^{d_2}$ and $D = g_1^\alpha$ during the dispute resolution.

Although Bao states that the problem of the fair payment is the result of the anonymity of the user, and also that due to the anonymity and untraceability of the system Wang's protocol cannot be improved to avoid the attack, the modification described in the previous section solves the vulnerability maintaining the anonymity and untraceability of the protocol, (the RCSS signature is maintained). Nobody, except the seller and the TTP, can verify the signed order agreement.

In section 4, we considered that two of the features of Wang's protocol were the reasons of the vulnerability. The first one, the relation between θ and the *e-coin* have been changed. The second one, the synchronous dispute resolution still remains. Without both of these features, the vulnerability has been removed.

6 Timeliness

Although the fair payment system has been improved to prevent Bao's attack, we have detected that the second cause, the lack of timeliness and the fact that the buyer cannot contact with the TTP can be the origin of a new problem. During the dispute resolution the buyer waits for a possible message from the TTP. If he doesn't receive any message from the TTP, it can be due to the lack of interest from M to finish the exchange (M has not sent the product) or due to a network failure (the message has been lost). U cannot resolve this situation because he cannot contact with the TTP. If the payment have been executed until step 4, that is, if U sends the answer to the challenge and M doesn't send the product to U (and doesn't contact with the TTP), then U cannot obtain the product, but even worse, he cannot use the coin again because with two answers to two different challenges he could be accused of double spending if in the future (there are not deadlines) M contacts with the TTP.

As a conclusion, the vulnerability described in [1] has been solved, however the system can be further improved to achieve timeliness. Recently, in TrustBus'05 the authors of [5] presented a timeliness fair exchange protocol that can be used with anonymous payment systems to achieve atomicity. This fair exchange protocol could be also used with Brand's payment system, avoiding the problem described above. In order to use [5] together with Brand's protocol, the payment system has to be slightly modified, like in section 3.2, to be adapted to the fair exchange protocol.

7 Conclusions

The paper analyses Bao's colluding attack applied to Wang's fair payment protocol and presents a modification to remove the vulnerability. The modified version adds an

element to link the e-coin and the signed order agreement that avoids the manipulation of the order agreement. This modification doesn't change the anonymity and untraceability of the protocol, the RCSS signature is maintained with this purpose. Wang's protocol includes a dispute resolution protocol where the seller can contact with a trusted third party to solve an unfinished exchange. However, the customer has to wait until the TTP contacts with him to solve the exchange. So the protocol doesn't achieve timeliness. This is the cause of a new problem of Wang's fair payment protocol, for this reason, a timeliness fair exchange is considered a better alternative.

References

1. Bao, F. "Colluding attacks to a payment protocol and two signature exchange schemes", ASIACRYPT 2004, LNCS 3329, pages 417-429, Springer Verlag, 2004.
2. Brands, S.: "Untraceable off-line cash in wallet with observers", Crypto'93, LNCS 773, pages 302-318, Springer Verlag, 1994.
3. Camp, J., Harkavy, M., Tygar, J.D. and Yee, B.: "Anonymous atomic transactions", 2nd USENIX workshop on electronic commerce, pages 123-133, 1996.
4. Jakobsson, M.: "Ripping coins for a fair exchange", Eurocrypt'95, LNCS 921, pages 220-230, Springer Verlag, 1995.
5. Payeras, M., Ferrer, J., Huguët, L: "Anonymous Payment in a Fair E-commerce Protocol With Verifiable TTP", Second International Conference on Trust, Privacy and Security in Digital Business, TrustBus'05. pages 60-69. LNCS 3592, Springer Verlag, 2005.
6. Schuldt, H., Popovivi, A. and Schek, H.: "Execution guarantees in electronic commerce payments.", 8th international workshop on foundations of models and languages for data and objects (TDD'99), LNCS 1773, Springer Verlag, 1999.
7. Tygar, J.D.: "Atomicity in electronic commerce", 15th annual ACM symposium on principles of distributed computing", pages 8-26, 1996.
8. Wang, C. H. "Untraceable fair network payment protocol with off-line TTP", ASIACRYPT 2003, LNCS 2894, pp. 173-187, Springer Verlag, 2003.
9. Xu, S., Yung, M., Zhang, G. and Zhu, H. "Money conservation via atomicity in fair off-line e-cash", International security workshop ISW'99, LNCS 1729, pages 14-31, Springer Verlag, 1999.

An Efficient Algorithm for Fingercodes-Based Biometric Identification

Hong-Wei Sun, Kwok-Yan Lam, Ming Gu, and Jia-Guang Sun

School of Software, Tsinghua University, Beijing 100084, PR China
sunhongwei@gmail.com, {lamky, guming, sunjg}@tsinghua.edu.cn

Abstract. With the emerging trend of incorporating biometrics information in e-financial and e-government systems arisen from international efforts in anti-money laundering and counter-terrorism, biometric identification is gaining increasing importance as a component in information security applications. Recently, fingercodes has been demonstrated to be an effective fingerprint biometric scheme, which can capture both local and global details in a fingerprint. In this paper, we formulate fingercodes identification as a vector quantization (VQ) problem, and propose an efficient algorithm for fingercodes-based biometric identification. Given a fingercodes of the user, the algorithm aims to efficiently find, among all fingercodes in the database of registered users, the one with minimum Euclidean distance from the user's fingercodes. Our algorithm is based on a new VQ technique which is designed to address the special needs of fingercodes identification. Experimental results on DB1 of FVC 2004 demonstrate that our algorithm can outperform the full search algorithm, the partial distance search algorithm and the 2-pixel-merging sum pyramid based search algorithm for fingercodes-based identification in terms of computation efficiency without sacrificing accuracy and storage.

Keywords: Biometric security, fingercodes; fingerprint matching; vector quantization.

1 Introduction

In the face of strong needs arisen from counter-terrorism and anti-money laundering requirements, new theories and technologies are being developed recently to address key issues related to the identification needs of financial and government systems. With the emerging trend of incorporating biometrics information in e-financial and e-government systems arisen from international efforts against terrorist financing and for effective border control, biometric identification is gaining increasing importance as a component in security applications.

Recently, fingercodes has been demonstrated to be an effective fingerprint biometric scheme [1]. Unlike minutiae-based matching algorithms, fingercodes captures both local and global details in a fingerprint; thus overcoming some of the major problems of minutiae-based matching. Fingercodes-based matching is based on the squared Euclidean distance between two corresponding fingercodes.

A fingerprint matching system may operate in one of the two modes: verification mode (1:1 matching) and identification mode (1: n matching). When

operating in the verification mode, it either accepts or rejects a user's claimed identity. Whereas, a fingerprint matching system operating in the identification mode establishes the identity of the user without a claimed identity information.

This research focus on the identification mode of fingercodes-based matching. The fingerprint matching system firstly computes the fingercodes of all registered users and stores them as fingerprint templates in the user database. The identification process then finds the minimum of the squared Euclidean distance between the user's fingercodes and all the templates in the database. By comparing this minimum distance with a predefined threshold, the system determines whether the user's fingercodes belongs to some registered user in the database. If so, it establishes the identity of the user based on the "nearest" template. However, fingercodes-based identification is a computation-intensive process as it needs to compute and compare the squared Euclidean distance between the user's fingercodes and all stored templates.

To enhance the performance of fingercodes-based identification, we formulate the identification process as a vector quantization (VQ) problem, and devise an efficient VQ algorithm to find the template with minimum Euclidean distance from the user's fingercodes. We observed that the fingercodes matching process is very similar to the encoding process in VQ [2,3], which is an efficient technique for low-bit-rate image compression. Many efficient algorithms have been developed to enhance the VQ encoding process [4,5,6]. These methods typically use the statistical features of a vector to estimate the Euclidean distance and reject most of the unlikely codewords without computing the actual Euclidean distances. Our algorithm also adopted this principle to accelerate the matching process for the fingercodes-based identification system.

Nevertheless, not all VQ techniques can be applied directly to fingercodes-based matching. VQ techniques designed for image compression use a small codebook, 512 codewords typically. Besides, the dimension of vectors representing image blocks is also small, typically 16 dimensions. Whereas, the dimension of fingercodes is of several hundred and the codebook size is determined by the number of registered users in the template database. Thus it is important to have a VQ technique which does not require storage space or pre-computed data set proportional to the dimension of the vectors or the codebook size.

This paper presents a new fingercodes identification algorithm based on VQ. The fingercodes biometric scheme is introduced in Section 2. In Section 3, we explain the fingercodes-based biometric identification and formulate the problem as a VQ encoding process. The new fingercodes identification algorithm will be described in detail in Section 4 which is followed by a presentation of the experimental results in Section 5. The paper is concluded in Section 6.

2 Fingercodes Biometric Scheme

Fingercodes has been demonstrated to be an effective fingerprint biometric scheme, which can capture both local and global details in a fingerprint [1]. The fingercodes-based matching algorithm uses a bank of Gabor filters to capture both

local and global details in a fingerprint as a fingercodes, which is represented by a fixed-length vector. The fingercodes scheme divides the fingerprint image of a user into a number of sectors, applies Gabor filters to transform the image in each sector; then obtains a real value from each sector by computing the average absolute deviation from the mean for each sector. This results in a k -dimensional real vector where k is the number of sectors on the fingerprint.

The fingercodes generation process [1] can be summarized by the following:

Step 1: Locate a reference point and determine the region of interest for the fingerprint image. The reference point of a fingerprint is defined as the point of maximum curvature of the concave ridges in the fingerprint image [7]. For fingercodes encoding, [1] located the reference point based on multi-resolution analysis of the orientation fields of the fingerprint image. The use of reference point helps address the translation invariance of fingercodes.

Step 2: Tessellate the region of interest around the reference point. The region of interest is divided into a series of B concentric bands and each band is subdivided into k sectors. In our experiments, we use four bands ($B = 4$) and each band is segmented into sixteen sectors ($k = 16$), thus resulting in a total of $16 \times 4 = 64$ sectors;

Step 3: Normalize the region of interest in each sector to a constant mean and variance. Let $I(x, y)$ denotes the gray value at pixel (x, y) , M_i and V_i , the mean and variance of the gray values in sector S_i , and $N_i(x, y)$, the normalized gray value at pixel (x, y) . For all the pixels in sector S_i , the normalized image is defined as:

$$N_i(x, y) = \begin{cases} M_0 + \sqrt{\frac{V_0 \times (I(x, y) - M_i)^2}{V_i}} & \text{if } I(x, y) > M_i \\ M_0 - \sqrt{\frac{V_0 \times (I(x, y) - M_i)^2}{V_i}} & \text{otherwise} \end{cases}$$

where M_0 and V_0 are the desired mean and variance values, respectively. The normalization process removes the effects of sensor noise and gray level deformation due to the finger pressure differences;

Step 4: Filter the region of interest in eight different directions using a bank of Gabor filters to produce a set of eight filtered images. Properly tuned Gabor filters can remove noise, preserve the true ridge and valley structures, and provide information contained in a particular orientation in the image. The typically used even symmetric Gabor filter [1] has the following general form in the spatial domain:

$$G(x, y; f, \theta) = \exp\left\{-\frac{1}{2}\left[\frac{x'^2}{\delta_x^2} + \frac{y'^2}{\delta_y^2}\right]\right\} \cos(2\pi f x')$$

$$x' = x \sin \theta + y \cos \theta$$

$$y' = x \cos \theta - y \sin \theta$$

where f is the frequency of the sinusoidal plane wave along the direction θ from x -axis, and δ_x and δ_y are the space constants of the Gaussian envelope along x and y axes, respectively.

Step 5: For each filtered output, compute the average absolute deviation from the mean (AAD) of gray values in individual sectors in filtered images to form the fingerCode. In our experiments, sixty-four features of each of the eight filtered images provide a total of 512 (64×8) features. Hence the fingercode of the user is represented by a collection of eight (8 filters) 64-dimensional real vectors.

The translation invariance of the fingercode is established by the reference point. To achieve approximate rotational invariance, the features in the fingercode are cyclically rotated. The fingercode is firstly rotated cyclically to generate five templates corresponding to five rotations ($0^\circ, \pm 22.5^\circ, \pm 45^\circ$) of the original fingerprint image. The original fingerprint image is then rotated by an angle of 11.25° and its fingercodes are generated by computing another five templates corresponding to five rotations. Thus, the database contains ten templates for each fingerprint.

To perform fingercode biometric verification, a user's fingercode is generated from his fingerprint image which is then matched against the stored fingercodes of the claimed identity. In this mode, fingercode matching is based on the Euclidean distance between the two corresponding fingercodes. The final matching distance score is taken as the minimum of the ten scores, i.e. matching of the input fingercode with each of the ten templates. This minimum score corresponds to the best alignment of the two fingerprints being matched.

3 Fingercode-Based Biometric Identification

When performing fingercode identification, a user supplies his fingerprint (query fingerprint) to the system without a claimed identity. The identification system then compares the supplied fingerprint with all the stored templates in order to determine the identity of the user. The fingerprint identification system thus firstly computes the fingercodes of all registered users and stores them as fingerprint templates in the database.

The fingercode identification process finds the minimum squared Euclidean distance between the fingercode of the query fingerprint and all the templates in the database. By comparing this distance with a predefined threshold, the system determines whether the query fingerprint matches some fingercode template in the database, and establishes the identity of the user.

Since the database stores ten templates for each fingerprint, if the full search (FS) algorithm is used, the matching process needs to compute the squared Euclidean distance between the fingercode of the query fingerprint and each of the templates in the database, thus it is computation-intensive.

To enhance the efficiency of fingercode-based biometric identification, we formulate the identification process as a VQ problem. In essence, the identification system needs to search through the fingercode database for the fingercode which is nearest (in terms of Euclidean distance) to the query fingerprint. This search problem is similar to the encoding process in VQ.

VQ is a mapping Q of a k -dimensional Euclidean space R^k into certain finite subset C of R^k , where C is the codebook with size N and each codeword $c_i = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$ in C is k -dimensional. The codeword searching problem in VQ is to assign one codeword to the input test vector such that the distance between this codeword and the test vector is the smallest among all codewords. Given one codeword $c_i = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$ and the test vector $x = \{x_1, x_2, \dots, x_k\}$, the squared Euclidean distance can be expressed as $d^2(x, c_i) = \sum_{j=1}^k (x_j - c_{ij})^2$.

In the fingercodeword identification system, since a fingercodeword typically has several hundred dimensions, and there are 10 templates for each fingerprint, the searching process is computation-intensive if the FS algorithm is used. From the above equation, each distance calculation needs k multiplication and $2k - 1$ additions. To encode the input vector, the FS method [2] computes the squared Euclidean distances between the input vector and each codeword and determines the best-matched one c_w by $d^2(x, c_w) = \min_{c_i \in C} d^2(c_i, x)$. To encode each input vector, the FS method needs N distance computations and $N - 1$ comparisons. In other words, it must perform kN multiplications, $(2k - 1)N$ additions and $N - 1$ comparisons, which is time-consuming.

In order to accelerate the VQ encoding process, many fast methods [4,5,6] have been proposed. These methods use an “estimate” of the Euclidean distance to quickly determine whether a codeword can be eliminated thus the actual Euclidean distance need not be computed. Suppose the running minimum for the input vector x is d_{min} , if the estimation for the Euclidean distance between x and current codeword c_i is larger than d_{min} and the corresponding real Euclidean distance is larger than the estimation, then we can safely reject c_i and avoid computing the actual Euclidean distance.

For example, the partial distance search (PDS) algorithm [4] uses the following rejection test $\sum_{j=1}^t (x_j - c_{ij})^2 \geq d_{min}^2$, for any $t < k$. If the partially calculated squared Euclidean distance from dimensions 1 to t is greater than the running minimum d_{min} , this codeword can be rejected without calculating the actual one in k dimensions.

As another example, the subvector (SV) method divides x and c_i into the first and second subvectors as $x_f = \{x_1, x_2, \dots, x_{k/2}\}$, $x_s = \{x_{k/2+1}, x_{k/2+2}, \dots, x_k\}$, $c_{i,f} = \{c_{i1}, c_{i2}, \dots, c_{ik/2}\}$, $c_{i,s} = \{c_{i(k/2+1)}, c_{i(k/2+2)}, \dots, c_{ik}\}$, respectively [5]. The sums, means and the variances of x and c_i are defined as $S_x = \sum_{j=1}^k x_j$, $S_{c_i} = \sum_{j=1}^k c_{ij}$, $M_x = S_x/k$, $M_{c_i} = S_{c_i}/k$, $V_x = \sqrt{\sum_{j=1}^k (x_j - M_x)^2}$, $V_{c_i} = \sqrt{\sum_{j=1}^k (c_{ij} - M_{c_i})^2}$. Similarly, the partial sums of each subvector are defined as $S_{x,f} = \sum_{j=1}^{k/2} x_j$, $S_{x,s} = \sum_{j=k/2+1}^k x_j$, $S_{c_i,f} = \sum_{j=1}^{k/2} c_{ij}$, $S_{c_i,s} = \sum_{j=k/2+1}^k c_{ij}$.

The SV method uses the following 3-step rejection test flow. If any inequality holds, it rejects c_i as the “nearest” codeword.

- Step 1.** $(S_x - S_{c_i})^2 \geq kd_{min}^2$.
- Step 2.** $(S_x - S_{c_i})^2 + k(V_x - V_{c_i})^2 \geq kd_{min}^2$.
- Step 3.** $(S_x - S_{c_i})^2 - 2(S_{x,f} - S_{c_i,f}) \times [(S_x - S_{c_i}) - (S_{x,f} - S_{c_i,f})] \geq (k/2)d_{min}^2$.

In order to realize recursive computation in a memory efficient way, 2PM SP as shown in Figure 1 was proposed in [6] for codeword search in VQ. A hierarchical rejection rule is set up as

$$d^2(x, c_i) = d_u^2(x, c_i) \geq \dots \geq 2^{-(u-v)} d_v^2(x, c_i) \\ \geq \dots \geq 2^{-(u-1)} d_1^2(x, c_i) \geq 2^{-u} d_0^2(x, c_i).$$

The squared Euclidean distance (i.e. the test function) at the v^{th} level of the hierarchy for $v \in [0, u]$ is $d_v^2(x, c_i) = \sum_{m=1}^{2^v} (S_{x,v,m} - S_{c_i,v,m})^2$ where $S_{x,v,m}$ is the m^{th} pixel at the v^{th} level for x and $S_{c_i,v,m}$ similarly defined for c_i . For a k -dimensional vector, $u = \log_2 k$.

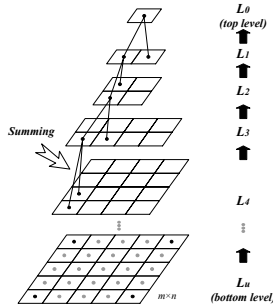


Fig. 1. A 2-pixel-merging sum pyramid

Thus, at any v^{th} level for $v \in [0, u]$, if the inequality $2^{-(u-v)} d_v^2(x, c_i) \geq d_{min}^2$ holds, then c_i can safely be rejected at the v^{th} level.

4 Efficient Fingerprint Identification Based on VQ

Due to the fundamental distinction between the fingerprint identification system and the image compression system, not all VQ techniques can be applied directly to fingerprint searching. The major differences include: (1) Codebook size in image compression is fixed, typically 512; whereas, fingerprint stored in database is very large and depends on the number of registered users in the system. (2) Image vector dimension is small, typically 16; whereas fingerprint vector dimension is much larger, e.g. $4 \times 16 \times 8$ in our examples. Thus a good fingerprint identification system should adopt a VQ technique which does not require storage space or pre-computed data set proportional to the dimension of the vectors or size of the user population.

Based on the above analysis, we firstly investigate some suitable fast VQ encoding methods and evaluate their performance for fingerprint identification. In this study, we use benchmarking data set DB1 from FVC 2004 [8] for our experiments. There are a total of 100 fingers and 8 impressions per finger (800

impressions) in this database. We use the first impression of each finger to test the performance of the PDS method [4], the SV method [5] and the 2PM SP method [6]. We then use the other 700 impressions as the registered fingerprints, compute 10 fingercodes for each impression and store them as the templates, which results in $700 \times 10 = 7000$ templates.

In the first experiment (Fig. 2), we used the first 1000 templates in the database to simulate a fingerprint identification system with a small population. In the second experiment (Fig. 3), we used all the 7000 templates in the database, which simulate a fingerprint recognition system with a larger population.

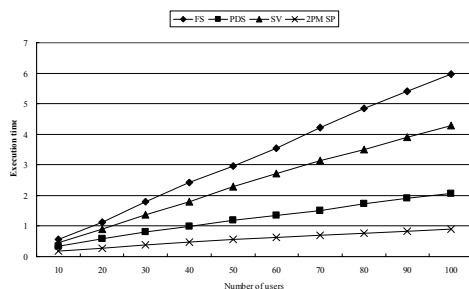


Fig. 2. Performance comparison for small number of users

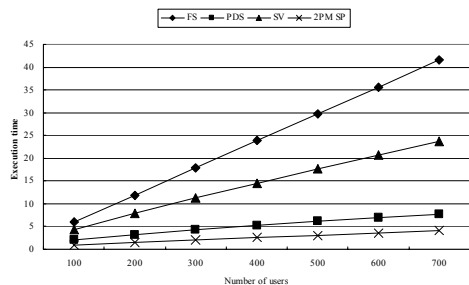


Fig. 3. Performance comparison for large number of users

From Figures 2 and 3, it is clear that the PDS and 2PM SP methods are more efficient and scalable for fingercodes identification. Note that both the PDS and 2PM SP schemes do not require additional storage space proportional to the user population size or vector dimensionality.

Our new scheme is based on and enhanced from these two methods. We propose an efficient search algorithm to find the minimum squared Euclidean distance, which combines the PDS algorithm [4] and the 2PM SP algorithm [6]. Since there will be more levels for fingercodes than the vectors in VQ and using the first several levels can hardly reject enough fingercodes, so we propose

a new scheme, namely the truncated 2PM SP, for the fingercode identification system. To achieve the improvement, the truncated 2PM SP scheme begins the computation of the squared Euclidean distance at the s^{th} level instead of the 0^{th} level, and at the s^{th} level, we introduce PDS algorithm. The truncated 2PM SP algorithm consists of the following five steps:

Step 1: Convert each template in the database to an $m \times n$ matrix as shown in Figure 4 ($m = 16, n = 32$ in our experiment).

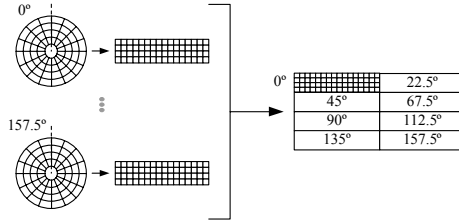


Fig. 4. Convert each template in the database to an $m \times n$ matrix

Step 2: Construct the truncated 2PM SP by computing the pyramid from the bottom level to the s^{th} level for each matrix computed in Step 1 ($s = 5$ in our experiment). If there are N users, and 10 templates for each user in the database, $10N$ truncated 2PM SP are constructed. That is, each template in the database is associated with one truncated 2PM SP.

Step 3: For the query fingercode x , the truncated 2PM SM of x is constructed. Select the first template c_1 in the database to be the current template “closest” to the query fingercode x . Compute and temporarily store the running minimum $d_{min}^2 = d^2(x, c_1)$ and $d_{v,min}^2 = 2^{(u-v)}d_{min}^2$ for $v \in [s, u]$.

Step 4: For any other template c_i in the database, execute the rejection tests from the s^{th} level to the bottom level. At the s^{th} level, if the partially calculated squared Euclidean distance is greater than the running minimum $d_{s,min}^2$, this template can be rejected. If c_i goes through all the rejection tests and arrives the bottom level, then recompute and store the running minimum $d_{min}^2 = d^2(x, c_i)$ and $d_{v,min}^2 = 2^{(u-v)}d_{min}^2$ for $v \in [s, u]$.

Step 5: If there is no more candidate template, the current “so far” best-matched template is the real best-matched one, and if d_{min}^2 is less than the predefined threshold, the index of this template is used to establish the identity of the user.

5 Experimental Results

Our algorithm is compared with the PDS and 2PM SP algorithms through experiments. In our experiments, we use the same parameters as in Section 4. The improvement ratio in terms of execution time required in the proposed algorithm

Table 1. Comparison of average online execution time per query fingercodes

Number of users	Number of templates	T_{PDS} (s)	T_{2PM} (s)	T_{T2PM} (s)	R_1 %	R_2 %
10	100	0.345	0.178	0.143	58.5	19.4
20	200	0.578	0.277	0.213	63.2	23.1
30	300	0.809	0.381	0.291	64.1	23.8
40	400	0.998	0.466	0.349	65.1	25.1
50	500	1.187	0.551	0.408	65.6	25.9
60	600	1.344	0.626	0.457	66.0	27.0
70	700	1.500	0.697	0.503	66.4	27.7
80	800	1.717	0.764	0.549	68.0	28.1
90	900	1.898	0.826	0.588	69.0	28.9
100	1000	2.073	0.888	0.625	69.8	29.6
200	2000	3.195	1.470	0.980	69.3	33.4
300	3000	4.363	2.023	1.319	69.8	34.8
400	4000	5.217	2.545	1.642	68.5	35.5
500	5000	6.204	3.063	1.963	68.4	35.9
600	6000	7.006	3.570	2.281	67.4	36.1
700	7000	7.787	4.080	2.597	66.6	36.4

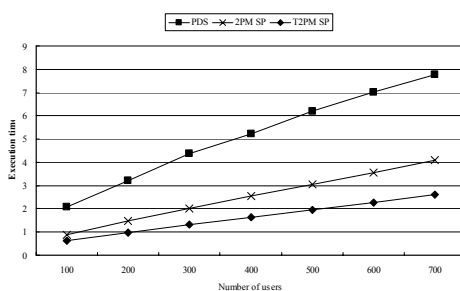


Fig. 5. Performance comparison for large user population

over the PDS and 2PM SP algorithms are denoted by $R_1 = \frac{T_{PDS} - T_{T2PM}}{T_{PDS}} \times 100$ and $R_2 = \frac{T_{2PM} - T_{T2PM}}{T_{2PM}} \times 100$ respectively. Here T_{PDS} , T_{2PM} and T_{T2PM} denote the average online execution time of the 100 tested fingercodes required in the PDS algorithm, the 2PM SP algorithm and the proposed algorithm, respectively.

Table 1 illustrates the respective performance comparisons where the time unit is denoted by ‘s’. As can be seen from Table 1, our algorithm performs more efficiently than both the PDS algorithm and the 2PM SP algorithm. More importantly, experimental results also illustrate the scalability of our algorithm since our algorithm performs well even with large user population as shown in Figure 5. Furthermore, our algorithm outperforms the FS algorithm, the PDS algorithm and the 2PM SP based search algorithm for fingercodes identification in terms of computation efficiency and yet without sacrificing accuracy and storage. (2PM SP requires k memories for a k -dimensional vector [6]).

6 Conclusions

In this paper, we proposed an efficient fingercodes identification technique based on an improved VQ encoding algorithm. The new algorithm finds the minimum squared Euclidean distance by combining the partial distance search and the

2-pixel-merging sum pyramid in VQ encoding. Because of the test structure of the truncated 2PM SP algorithm, our algorithm is more efficient than both the PDS and basic 2PM SP algorithms.

Experimental results on DB1 of FVC 2004 [8] demonstrated that our algorithm can outperform the full search algorithm, the partial distance search algorithm and the 2-pixel-merging sum pyramid based search algorithm for fingerprint-based fingerprint matching in terms of computation efficiency without sacrificing accuracy and storage.

Acknowledgement. This research was partly funded by the Chinese National Science Foundation (Project Number: 90412007), P. R. China and the Biometric Cryptography project of PrivyLink Pte Ltd (Singapore).

References

1. A.K. Jain, S. Prabhakar, L. Hong, S. Pankanti. "Filterbank-based fingerprint matching", *IEEE Trans on Image Processing*, 2000, 9(5): 846-859.
2. Y. Linde, A. Buzo, R. M. Gray. "An algorithm for vector quantizer design", *IEEE Trans on Comm*, 1980, 28(1): 84-95.
3. N.M. Nasrabadi, R.A. King. "Image coding using vector quantization: a review", *IEEE Trans on Comm*, 1988, 36(8): 957-971.
4. C. Bei, R.M. Gray. "An Improvement of the Minimum Distortion Encoding Algorithm for Vector Quantization", *IEEE Trans on Comm*, 1985, 33(10): 1132-1133.
5. Z. Pan, K. Kotani, T. Ohmi. "Improved fast encoding method for vector quantization based on subvector technique", *2005 IEEE International Symposium on Circuits and Systems*, 2005: 6332-6335.
6. Z. Pan, K. Kotani, T. Ohmi. "A memory-efficient fast encoding method for vector quantization using 2-pixel-merging sum pyramid", *2004 IEEE Int'l Conf on Acoustics, Speech and Signal Processing*, 2004, 3: 669-672.
7. X.D. Jiang, M.H. Liu, A.C.C. Kot. "Reference Point Detection for Fingerprint Recognition", *the 17th Int'l Conf on Pattern Recognition*, 2004, 1: 540-543.
8. FVC 2004. "The 3rd International Fingerprint Verification Competition", <http://bias.csr.unibo.it/fvc2004/>.

Robustness of Biometric Gait Authentication Against Impersonation Attack

Davrondzhon Gafurov, Einar Snekkenes, and Tor Erik Buvarp

Norwegian Information Security Lab
Department of Computer Science and Media Technology
Gjovik University College
P.O. Box 191 2802 Gjovik, Norway
{davrondzhon.gafurov, einar.snekkenes, tor.buvarp}@hig.no

Abstract. This paper presents a gait authentication based on time-normalized gait cycles. Unlike most of the previous works in gait recognition, using machine vision techniques, in our approach gait patterns are obtained from a physical sensor attached to the hip. Acceleration in 3 directions: up-down, forward-backward and sideways of the hip movement, which is obtained by the sensor, is used for authentication. Furthermore, we also present a study on the security strength of gait biometric against imitating or mimicking attacks, which has not been addressed in biometric gait recognition so far.

1 Introduction

Automatic biometric authentication is the process of verifying the claimed identity of individual by his or her physiological or behavioral characteristics. Examples of human traits that can be used for automatic biometric authentication include fingerprint, iris, retina, face, voice, handwriting, gait¹, etc. Gait biometric has an advantage of being non-intrusive and the ability to be captured from the distance when other type of biometrics are not available. Earlier studies on gait recognition showed promising results, usually with small sample size [1,2,3,4,5,6,7]. E.g. with the database of 16 gait samples from 4 subjects and 42 gait samples from 6 subjects Hayfron-Acquah [1] achieved correct classification rates of 100% and 97%, respectively. However, recent studies with a larger sample size confirm gait as having discriminating power from which individuals can be identified [8,9,10]. Most of the work done in the direction of gait recognition uses machine vision techniques to extract gait patterns from video or image sequences [1,2,3,4,5,6,7,8,9,10].

In this paper we present a gait authentication based on time-normalized gait cycles. Unlike most of the previous works in gait recognition, which extract gait patterns using machine vision techniques, in our approach gait patterns are obtained from a MEMS device (physical sensor) attached to the hip. MEMS (Micro-Electro-Mechanical System) is an integration of mechanical elements,

¹ Gait is a person's manner of walking.

sensors and electronics on a common framework [11]. Acceleration in 3 directions, up-down, forward-backward and sideways of the hip movement, which is obtained by the MEMS device, is used for authentication. Despite much research work being carried out in gait recognition, however to our knowledge, no work has been reported that investigates the possibility of spoofing gait biometric. Furthermore, this paper presents a study on the security strength of gait biometric, particularly its robustness against imitating or mimicking attacks. The rest of the paper is structured as follows: section 2 describes attacks and an evaluation scenario of biometric system, section 3 contains a description of the MEMS device, gait verification method, experiments and results, section 4 contains discussion, section 5 outlines some possible areas of application for MEMS-based gait authentication, and finally section 6 concludes the paper.

2 Attacks and Evaluation Scenarios for the Biometric System

2.1 Performance Evaluation Scenarios

There are two important types of submitting biometric sample to the authentication system. First is a *genuine attempt* which is a self verification attempt, when an individual submits his own biometric feature to match against his own template. The second one is a *non-genuine or impostor attempt* which is a non-self verification attempt when an individual submits his own biometric feature to match against another person's biometric in the template. We subdivide non-genuine attempts into the three following groups:

- *Passive impostor attempt* is an attempt when an individual submits his own biometric feature as if they were attempting successful verification against his own template but in fact is being compared against non-self template.
- *Active impostor attempt* is an attempt when an individual changes his biometric with the aim to match another targeted person, and verified against this targeted person's template.
- *Non-passive and non-active impostor attempt*. An example of this attempt is when an active impostor trial is compared against not the targeted person but someone else's template.

From these three subgroups of impostor attempts only the first two ones are important. Conventionally, performance of the biometric systems is evaluated under *friendly scenario*, meaning that all impostor attempts consists of only passive impostor trials. We define *hostile scenario evaluation* when biometric system's impostor trials consists of active impostor attempts. In general, friendly scenario evaluation related to the discriminating power of biometric, whereas the hostile scenario shows the performance of the system against attacks.

2.2 Attacks on Biometric System

Typical points of attack on biometric authentication system, defined by Ratha et al.[12], are depicted in Figure 1.

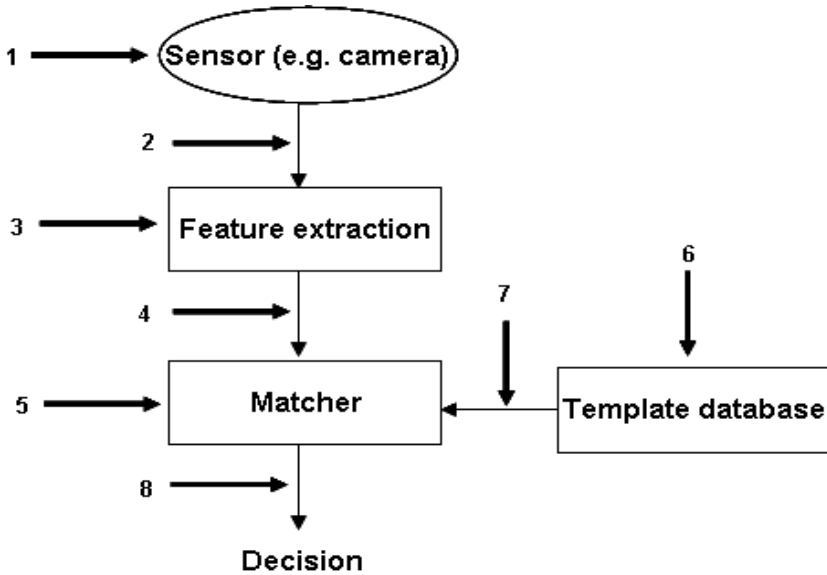


Fig. 1. Eight possible attack points in a biometric authentication system

1. First attack is presenting a fake or imitated biometric to the sensor.
2. Second type of attack involves submitting of previously obtained digital biometric signal.
3. In third type of attack, feature extractor is attacked so that it produce feature values dictated by attacker.
4. Extracted feature values are substituted by the ones selected by the attacker in the fourth type of attack.
5. In fifth type of attack, the score of matcher changed to produce desired high or low matching score.
6. An attack on database of biometric templates forms the sixth type of attack.
7. The seventh attack targets the transmission channel between template database and matcher module.
8. Last type of attack involves alternation of decision (accept or reject).

In this work we study the possibility of imitating or mimicking another person's walking manner, which is related to attack type 1. We only consider minimal effort attacks. By minimal effort attack we mean those type of attacks that do not require deep knowledge and experience of the system.

3 Gait Authentication Technology and Results

3.1 MEMS Device and Feature Vector

The MEMS device used to collect gait data resembles a memory stick device and has following main features: storage capacity (64-256+MB), USB and wireless

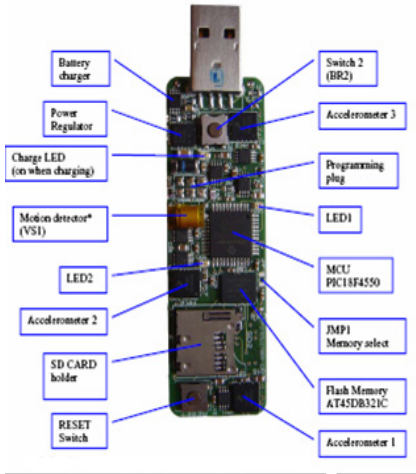


Fig. 2. MEMS device used for collect- ing acceleration data



Fig. 3. MEMS device attached to the hip

Bluetooth interfaces for data transfer, and 3 acceleration sensors, see Figure 2 and Figure 3. It records acceleration at the rate of about 100 samples per second.

From the output of device acceleration signals in three directions: vertical X , backward-forward Y , and sideway Z are obtained. However, instead of analyzing these raw acceleration signals separately, which might be sensitive to the device’s placement and orientation, we use invariant combination of them, so called resultant gait signal, which is calculated as follow:

$$R_i = \sqrt{X_i^2 + Y_i^2 + Z_i^2}, i = 1, \dots, k$$

where R_i is the resultant acceleration at time i , X_i , Y_i and Z_i are vertical, forward-backward and sideway acceleration at time i , respectively, and k is the number of recorded samples.

3.2 The Cycle Length Method

The resultant gait signals were compared based on time-normalized cycle length method. First, from 3 acceleration signals resultant gait signal is computed. The intervals between samples are not the equal, so in the second step resultant gait signal is interpolated. To reduce the level of noise in the signal a moving average (MA) filter is applied. Then, cycles are detected, normalized and averaged. Finally, Euclidean distance between two averaged cycles is computed as follow,

$$dist(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2},$$

where $A = (a_1, \dots, a_n)$ and $B = (b_1, \dots, b_n)$ are two averaged cycles (feature vectors), and a_i and b_i are resultant acceleration values at time point i . This distance value represents similarity score of two resultant gait signals. Ideally, for genuine trials the similarity scores should be smaller than for impostor trials. The steps involved in comparing gait signals are visualized in Figure 4.

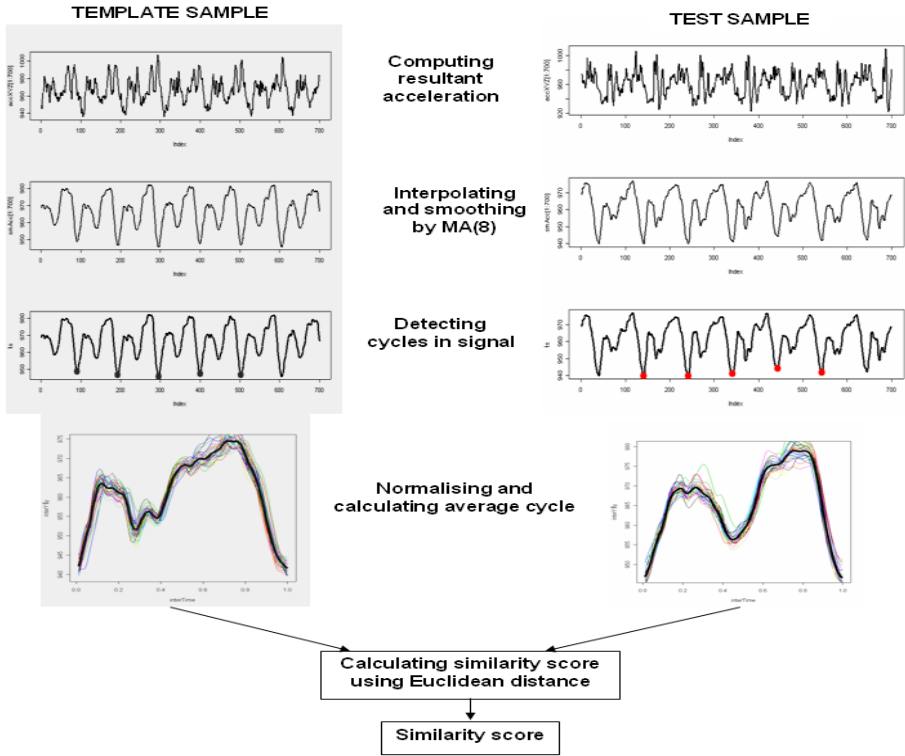


Fig. 4. Steps for comparing gait samples

3.3 The Data Set

Friendly scenario: In the friendly scenario 22 subjects are participated, 5 female and 17 male in the age range 20-38. Subjects were asked to walk normally on a level floor. MEMS device was attached to the hip of the subjects as shown in Figure 3. Subjects walked 3 rounds and each time device was removed and re-attached again to simulate realistic environment. Each walking round was splitted into two, thus we obtain 6 gait samples per subject (6 data sets). Every gait sequence represents about 20 seconds of walking. We conducted cross comparison among 6 data sets, in this way we generated 330 genuine and 6930 passive impostor trials.

Hostile scenario: In this scenario 20 subjects from the friendly scenario participated. Every subject was paired with another one. Everyone was told to study

his pairs' walking style and try to imitate him or her. One subject from the pair acted as an attacker, the other one as a target, and then the roles were exchanged. Pairing was conducted randomly, not on the similarity of physical characteristics, i.e. attacker and target should not necessarily be of similar height or weight. Everyone made 2 rounds of mimicking. In first attempt target person was walking in front of attacker, and in the second one attacker was mimicking alone. Similarly, every round was splitted into two samples, thus we have 4 imitated gait samples per attacker. All attackers were amateurs and they study target person only visually. The only information about the gait authentication system they knew was that acceleration of normal walk is used. Imitated gait samples were compared only with the targeted subject's ones, thus producing 480 active impostor trials.

Both experiments (friendly and hostile) were conducted in the same indoors location.

3.4 Results

The performance of the system in a friendly scenario in terms of decision error trade-off curve (DET) is shown in Figure 5. The DET curve is a plot of false accept rate (FAR) versus false reject rate (FRR), and it characterizes performance of the biometric authentication system under different operational points (thresholds) [13]. An interesting point in the DET curve is the EER (equal error rate) where FAR=FRR. EER of our method is about 16%, which means that out of 330 genuine attempts 53 are wrongfully rejected, while out of 6930 passive impostor attempts 1109 are wrongfully accepted.

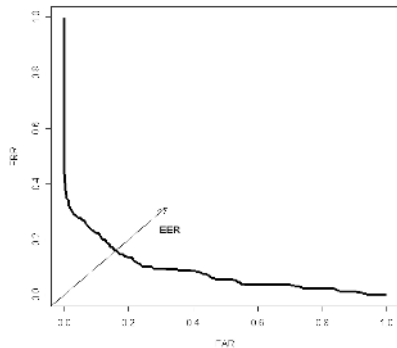


Fig. 5. Performance of the resultant gait signal in terms of the DET curve

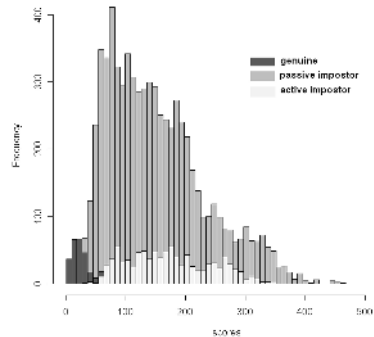


Fig. 6. Distributions of genuine, passive and active impostor scores

For the analysis of the hostile scenario we apply statistical techniques to check the difference between passive impostor trials and active impostor trials. The distributions of the genuine, passive impostor and active impostor are given in Figure 6. Different statistics of passive and active impostor trials are summarized

in Table 1. D-prime value represents separability of two normal distributions [14]. In our case it shows separability between genuine and passive impostor distributions, and between genuine and active impostor distributions. The larger the value of d-prime the more separable the two distributions are. It appears that genuine and active impostor distributions are more separable than genuine and passive impostor distributions. For comparing passive and active distributions we have stated a hypothesis that the mean of passive impostor is greater or equal to the mean of active impostor, and alternative is that active impostor’s mean is greater than passive impostor’s mean (attacks are not successful), i.e.

$$H_0 : \mu_{active} \leq \mu_{passive}$$

$$H_A : \mu_{active} > \mu_{passive}$$

To evaluate this hypothesis we applied both parametric and non-parametric tests. First, assuming normality we applied t-test and found little evidence towards null-hypothesis ($p\text{-value} = 0.0005$). Second test we applied was nonparametric Wilcoxon (or Mann-Whitney) test. Similarly, we have found little evidence to support the hull hypothesis in this test, either ($p\text{-value} = 0.000004$).

Table 1. Statistics on active and passive impostor distributions

Statistics	Active impostor	Passive impostor
Size	480	6930
Mean	163	152.3
Standard deviation	67.8	80.6
Standard error	0.97	3.1
Skewness	0.32	0.82
D-prime	1.415	1.142

4 Discussion

The performance of the method in a friendly scenario is comparable with other methods [15,16,17]. MEMS based gait authentication is very recent topic in biometric gait recognition. To our knowledge, so far only 3 works have been reported in this direction [15,16,18]. Our approach is similar to the work by Ailisto et al. [15] and Mantyjarvi et al. [16]. However, in [15,16] acceleration in side-way direction is not taken into account, and their MEMS device records acceleration at the higher rate (256 samples/sec.). In [18] acceleration of the lower leg is investigated for authentication, and MEMS device used records acceleration at the rate of 16 samples per second. Operational mode of the MEMS based gait systems are usually different. MEMS-based systems operate on verification, or sometimes also called authentication, mode (one to one comparison), whereas vision-based system on identification mode (one to many comparisons).

Unlike, for example voice [19,20] or signature [21,22] biometrics, for which impersonation attacks (or mimicking) have been studied, security of gait biometric has not received attention. In this work we also studied mimicking of someone’s

else gait in terms of the differences between passive and active impostor distributions. Our preliminary analysis suggests that this type of attack might not be successful for gait biometric. Gait of humans is a complex process that involves nervous and musculo-skeletal system [23]. When a person was told to walk as a someone else we believe he or she is given a restriction, since he has to walk other than his normal habituated gait. Because of this he or she may fail to produce gait patterns exactly as a target person. Another reason of the failure might be due to the difference on physical characteristics between attacker and target. Therefore, we hypothesize that in general, for gait biometric probability of success in an active impostor attempt might not be greater than probability of success in a passive impostor attempt, provided minimal effort attacks. We emphasize minimal effort since it is not known yet whether training can increase success of active impostor attempts. In other words, when person is trained can he or she walk as another person.

5 Application

MEMS-based gait recognition lacks some difficulties of vision based systems, such as background subtraction, viewing angle, lighting conditions etc. However, it shares common factors that can alter gait of the person like surface, injury, carrying load and so on. Applications of vision-based gait system generally focus on forensics and surveillance, while MEMS-based system on authentication and access control. MEMS-based gait system has been proposed for protecting mobile and portable devices [15,16,18]. For example, indeed it might be preferable to make few steps instead of recalling rarely used password for activating mobile phones. Another interesting application for MEMS-based gait authentication system can be in the area of wearable computing. Wearable computers are computers that can be worn effortlessly, run continuously and be operated hands-free [24]. The issues of unobtrusiveness and user's attention are important in such computing environment [25]. Therefore gait can be a good candidate for authentication to wearable devices, provided that error rates can achieve satisfactory levels. However, still due to higher error rates gait cannot be considered as a strong authenticator, nevertheless, when combined with other types of authentication (e.g. using more than one device) it may improve security and enhance usability of the system.

6 Conclusion and Future Work

In this paper we provided another evidence towards MEMS-based gait authentication by showing possibility of using hip acceleration for authentication. Using gait samples from 22 subjects we achieved EER of 16%. However, development of better algorithms is required for lowering error rates. We also outlined some possible application areas for MEMS-based gait recognition system. In addition, we addressed the topic of mimicking gait, and our preliminary analysis suggests that minimal effort impersonation attacks on gait biometric might not be successful. However, further studies with larger sample size are necessary in this

direction. It also is important to study whether it is possible to improve impersonation attacks by training. Another topic of interest would be to see which of the attackers are relatively more successful on mimicking, and secondly which of the targets are relatively easy to imitate. Due to the fact that MEMS based gait authentication is very recent, there is no established database, which would allow us to compare performance of different algorithms under common bases. All these topics will constitute basis for our future work.

References

1. James B. Hayfron-Acquah, Mark S. Nixon, and John N. Carter. Automatic gait recognition by symmetry analysis. In *Audio- and Video-Based Biometric Person Authentication*, pages 272–277, 2001.
2. J.Shutler and M.Nixon. Zernike velocity moments for description and recognition of moving shapes. In *British Machine Vision Conference*, pages 11/1 – 11/4, 2001. Session 8: Modelling Behaviour.
3. D. Cunado, M. Nixon, and J. Carter. Automatic extraction and description of human gait models for recognition purposes. In *Computer Vision and Image Understanding*, pages 1 – 41, 2003.
4. Liang Wang, Weiming Hu, and Tieniu Tan. A new attempt to gait-based human identification. In *International Conference on Pattern Recognition*, pages 115–118, 2002.
5. C. BenAbdelkader, R. Cutler, and L. Davis. Stride and cadence as a biometric in automatic person identification and verification. In *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 357–362, May 2002.
6. C. BenAbdelkader, R. Cutler, H. Nanda, and L. Davis. Eigengait: Motion-based recognition of people using image self-similarity. In *Audio- and Video-Based Biometric Person Authentication*, 2001.
7. Amos Y. Johnson and Aaron F. Bobick. A multi-view method for gait recognition using static body parameters. In *Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 301–311, June 2001.
8. Sudeep Sarkar, P. Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(2):162–177, 2005.
9. Toby H. W. Lam and Raymond S. T. Lee. A new representation for human gait recognition: Motion silhouettes image (msi). In *ICB*, pages 612–618, 2006.
10. Yuan Wang, Shiqi Yu, Yunhong Wang, and Tieniu Tan. Gait recognition based on fusion of multi-view gait sequences. In *ICB*, pages 605–611, 2006.
11. Jeremy A. Walraven. Introduction to applications and industries for microelectromechanical systems (MEMS). In *International Test Conference*, pages 674–680, 2003.
12. Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. An analysis of minutiae matching strength. In *Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 223–228, June 2001.
13. A.Martin, G.Doddington, T.Kamm, M.Ordowski, and M.Przybocki. The DET curve in assessment of detection task performance. In *Eurospeech'97*, pages 1895–1898, 1997.

14. Bolle R.M., Pankanti S., and Ratha N.K. Evaluation techniques for biometrics-based authentication systems (FRR). In *15th International Conference on Pattern Recognition*, pages 831 – 837, September 2000.
15. Heikki J. Ailisto, Mikko Lindholm, Jani Mantyjarvi, Elena Vildjiounaite, and Satu-Marja Makela. Identifying people from gait pattern with accelerometers. In *Proceedings of SPIE Volume: 5779; Biometric Technology for Human Identification II*, pages 7–14, March 2005.
16. Jani Mantyjarvi, Mikko Lindholm, Elena Vildjiounaite, Satu-Marja Makela, and Heikki J. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005.
17. A.F. Bobick and A.Y. Johnson. Gait recognition using static, activity-specific parameters. In *Proceedings of the 2001 IEEE Computer Computer Vision and Pattern Recognition*, pages I-423 – I-430, 2001. vol.1.
18. Davrondzhon Gafurov, Kirsi Helkala, and Torkjel Sondrol. Gait recognition using acceleration from MEMS. In *1st IEEE International Conference on Availability, Reliability and Security (ARES)*, Vienna, Austria, April 2006.
19. Lindberg J. and Blomberg M. Vulnerability in speaker verification - a study of technical impostor techniques. In *Eurospeech*, pages 1211–1214, 1999.
20. Yee Wah Lau, Wagner M., and Tran D. Vulnerability of speaker verification to voice mimicking. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 145 – 148, October 2004.
21. Guo J.K., Doermann D., and Rosenfield A. Off-line skilled forgery detection using stroke and sub-stroke properties. In *15th International Conference on Pattern Recognition*, pages 355 – 358, September 2000.
22. Sung-Hyuk Cha and Tappert C.C. Automatic detection of handwriting forgery. In *Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 264 – 267, August 2002.
23. Christopher Vaughan, Brian Davis, and Jeremy O’Cononor. *Dynamics of human gait*. Kiboho Publishers, 1999.
24. Polly Huang. Promoting Wearable Computing: A Survey and Future Agenda. In *In Proc. of International Conference on Information Society in The 21st Century: Emerging Technologies and New Challenges*, November 2000.
25. Starner T.E. Attention, memory, and wearable interfaces. *Pervasive Computing, IEEE*, Vol. 1, Issue 4:88–91, Oct.-Dec. 2002.

From Features Extraction to Strong Security in Mobile Environment: A New Hybrid System

Stéphane Cauchie, Thierry Brouard, and Hubert Cardot

Computer Science Lab., University of Tours, 64 av. J.Portalis,
37200 Tours, France

{stephane.cauchie, thierry.brouard, hubert.cardot}@univ-tours.fr
<http://www.li.univ-tours.fr/>

Abstract. This article describes a complete original biometric system named Tactim based on the skin reaction coupled with security concepts. Even if the experiments showed that there is much work left, our approach has many advantages. Security and biometry are strongly linked together, the biometric features are original and strongly depend of the user's will, it can be easily interfaced to existing systems, so we finally propose a way to integrate this system in a Service Oriented Architecture and analyze its advantages and drawbacks.

1 Introduction

Security and biometry are independently well tried. Whereas biometry performs authentication or identification based on physical or behavioral characteristics, secure systems are designed so as to offer properties such as: Integrity, confidentiality, non-repudiation...

The last decade of research tends to bring together these two antagonist domains. Why such a choice? The objectives of security are ensured by authentication based on a logical entity (password, PIN, certificates). On the other hand biometry performs authentication based on physical features collected from the user himself. The hybridization of security and biometry leads to the building of a more powerful scheme. In this way we hope to take advantage of both methods and to overcome their inherent drawbacks.

[11] deals about four biometric issues: (i) Any system needs biometric data to process, (ii) Biometric features can not be changed, (iii) The decision layer gives a binary response and (iv) Biometric systems are sensitive to replay attacks. To overcome the exposed problems, the research focuses on two main topics.

The first one is the convergence of ideas surrounding the protection of the biometric data. From naive approaches such as (use of cryptographic protocols), it has appeared that, in an authentication scheme, centralized data must be prohibited. Then the combination with tokens opens new opportunities and precises the formulation of the problem: the stored data must be useless without the biometric features.

The decision layer is the second critical topic. The greatest risk is strongly associated to a binary (y/n) one. One can build a masquerading system.

Considering these actual works, we propose an architecture which respects the previously exposed constraints; (i) centralized data are prohibited, (ii) stored data are public, (iii) the decision layer is enforced: We conceal personal data in a mobile architecture (a smart card) via the Tactim biometric authentication which is our original type of biometry ([29]). The idea is to enclose the user's personal data within the smart card protected through a PIN (Personal Identification Number) authentication which is unbeknownst to him. The PIN is reconstructed from the user's acquisition and the publicly stored biometric data.

The presented solution showing a complete system (it can be used in any scenario), we think about its integration as an authentication service for Service Oriented Architecture (SOA). We propose a basic scheme to provide consistent authentication (abstraction of atomic process: Acquisition, authentication...).

The paper is organized as follows : We first make an overview of issues raised by the hybridization between cryptographic and biometric methods (section 2). Then, our proposed architecture is presented in section 3. The next section (4th) focuses on the pattern recognition problem related to our approach, and shows that Tactim's system is more reliable than fingerprints-based biometric systems. The integration of our system into a SOA is described in section 5, before the conclusion given in section 6.

2 Related Work

The most practical way of protecting the biometric data is the use of an adapted hardware. Through a standard storage device the solution proposed in [1,2] consists in the encryption of the database. Mutual authentication between components ensures the legitimate usability of the information. But using cryptographic tools to ensure the protection of the biometric system can not be considered as a real hybridization but more as a sequential approach. The use of tokens, or mobile device storage such as smartcard or javacard, are a popular alternative for authentication process. As smartcards are non-invasive devices, it can safely store personal information (such as passwords, certificates); several solutions currently propose the storage of the biometric data in such manner [4,5,6]. This implies that the global system knows the user's PIN which represents a potential risk. Existing systems can perform the decision within the device [7] but this kind of architecture does not properly solve the problem as when the access is granted, the biometric data can be retrieved. Nevertheless these approaches propose a mobile device and, at the same time, forbid centralized data, to eventually increase the difficulty in collecting the biometric data. Recently, a new approach was exploited: rendering the stored data useless. The BioHashing and FaceHashing methods exploit this objective: they are able to protect the data and let them be public [8,9,10]. A pseudo random generator (PRG) is stored in a token; the acquired biometric signal is mixed with the random stream before being presented to the decision layer. Any supplementary data stored are dependent of the biometric signal and the PRG, and while attackers can not have both the security is ensured.

The first security considerations for biometric process are done with the help of cryptographic tools [1,2]. Each component has to prove its identity to the others, and so the basic decision is entrusted by this mechanism. Cryptographic keys are commonly deduced from pseudo random numbers: [12] proposes to use directly the biometric features as a part of the pseudo random number generator seed. This approach does not take into account the noisy data property carried by the biometric signals.

The following models try to be noise tolerant. In [13,14], the key is hidden in a look-up table (LUT). The key can be recovered from the features; each feature gives a logical address in the LUT (according to a predetermined threshold). To generate a strong key, redundant information is used, thus decreasing the global entropy. The main problems are that the lack of entropy may conduct to a cryptanalysis of the generation space of the keys and that the determination of the thresholds leads to poorer results (the characteristics are taken into account independently).

Today, the most popular methods are based on the Corrected Byte Code Theory (CBCT). The goal of such methods is to be enable the correction of errors which might occur during an electronic transmission (from satellite, CD-ROM) [15]. The concept is easily transposed. Users own their biometrics features, which are transmitted to the biometric system (during the acquisition). The use of CBCT allows the correction of these errors in order to recover the initial data.

Using directly the CBCT on biometric system [16], prevents the changing of the system key. The use of additional data, combined with the result of the CBCT, provides a more secure scheme [17,18].

In [19] the authors proposed a new scheme based on the Biohashing method and the CBCT. Whereas the first method let the data be useless, the CBCT can recover predefined data. These data are used to initialize the Shamir secret-sharing algorithm.

Although these methods give acceptable results, the appearing problems are; the difficulty to adjust the False Accepted Rate in front of the False Rejected Rate [16].

3 The Proposed Architecture and Its Integration

The system is divided into three layers (see figure 1): Acquisition, storage and computing. A first device, the "sensor" takes care of acquisition and storage abilities. The sensor is dedicated to only one person. A second device, the "base", is shared and provides the computing. Doing like this gave us the compliance with the security recommendations. The acquisition is done by a sensor we have designed [29]. The storage is handled by well-known technologies like smartcards or javacards.

In a biometric scheme, during the enrolment step, the parameters of the system are adapted to the considered physical features of the user. Here the sensor and the base are linked. The user is invited to sign several times on his own sensor. With this set of signatures, biometric data is computed by the basis. Two

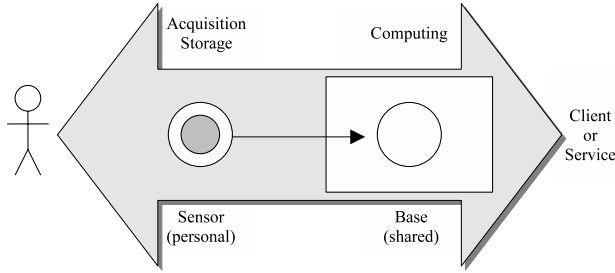


Fig. 1. Global representation of our architecture

steps are followed in order to ensure both public data storage and the PIN's secretive aspect. First, the biometric data is hashed by SHA-256 to produce the 8-bytes PIN and customize the user's smart card. Secondly, CBCT creates redundant information from it (C_I , corrective information). C_I is the biometric data publicly stored.

In this scheme, the PIN is unknown to anyone as to enforce security. To avoid the compromission, no sufficient information is given by the sole biometric data (ie: without the user's signature).

4 The Biometric Kernel

Biometry often needs pattern recognition tools to be really efficient. In this section we present these tools and how our hybridization was done.

4.1 Tactim's Biometric Features

The recent study (50 years) in derma-sciences details the composition of the human skin composition but also its interactions with the brain. The glabrous skin (without bristle) contains numerous mechanoreceptors. Each receptor reacts to a particular stimulus (like pressure, temperature, vibration,... for more details see [24]). Our purpose deals with Pacinian corpuscles (activated through vibration): a type of mechanoreceptors which can be found in the deep layer of the skin. The number of these receptors is the highest in the middle finger. When you rub this finger on a coarse plan, the vibration due to fingerprints innervates Pacinian corpuscles, producing a signal dispatched to the brain. The latter generates a feedback (each stimulus is followed by one) which modifies the skin's density and the muscular tonus. This modification allows us to acquire a signal, unique to each individual (due to the interdependence between action, stimulus and feedback).

4.2 User's Signature Computation

The biometric information is read by the sensor ([29]). The produced signal is digitalized (32 bits/sample, 44100 Hz) and then filtered. Having a mean duration of 0.8 second, this signal contains about 170 times the same pattern. This

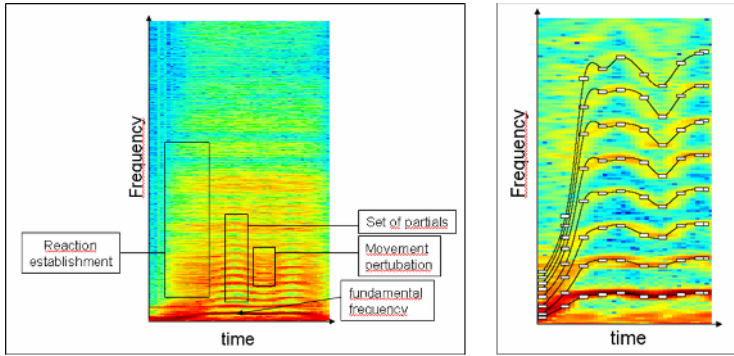


Fig. 2. Spectrogram and features extraction of a Tactim's biometric signal

pattern, distorted in amplitude, and has a period varying between 2 and 9 milliseconds (100-500 Hz). An example of the spectrogram of the signal is given in figure 2. Some areas are represented: a fundamental frequency, a set of partials, and a transition phase during which the physiological reaction appears.

Some well-known methods in speech recognition like cepstral coefficients [21] or dynamic time warping technique [22] have been tried without significant results. Then the tests showed that methods for tracking partials like MAQ or its derivatives [23] are not precise enough in this case. A specific algorithm using b-splines was designed to characterize the signal. This one tracks the partials to extract the features: C_1 the fundamental frequency, C_2 the mean interval between successive partials and C_3 the highest frequency having a measurable partial.

4.3 The Decision Step

The authentication can be viewed as a "one class classification problem". The taken decision represents the acceptance (the sample is recognized as belonging to the genuine user) or the reject (the only one alternative). The problem consists in the comparison of a given signature to test T and a set S of R reference signatures s_r , each one represented by a vector of N values denoted C_n^r . During the enrolment, the inner class variance E_n and the outer class variance A_n of the n -th feature were computed. Then a weighted-euclidian distance where the weight α_n of the n -th feature is given by $\alpha_n = \frac{E_n}{A_n}$ is used to compute the distance between T and each one of the R reference signatures. So, we can find D_{min} the lower distance and D_{max} the greater one. Given two acceptance thresholds $[H_{min}, H_{max}]$, three decision criteria were tested:

1. **DC1:** if $D_{min} \leq H_{min}$ we accept the current signature, we reject it otherwise.
2. **DC2:** if $D_{max} \leq H_{max}$ we accept the current signature, we reject it otherwise.
3. **DC3:** DC1 and DC2

4.4 Integration with Security

To generate the system key (for us the PIN), we use the popular Corrected Byte Code Theory [15] created for electronic transmission issues and this, following two steps. First, before an emission, corrective information (Ic) is computed then appended to the message T to obtain T' : $T' = \{T, Ic\}$. Secondly, after the emission, the transmitted message may be altered by errors (ϵ). So $(T' + \epsilon)$ can be corrected using Ic in order to correct T' and then recover T . Applied to an authentication scheme, the training phase consists in finding the biometric template needed to compute Ic . During the authentication, the biometric system receives an altered version of this template. As above, the CBCT is used to reconstruct T . In our system, we consider that a hashed version of T represents our system key.

The aim is to recover the template only if the features are provided by a genuine user. It is obvious that the length of Ic is directly link to the correction capacity and so, to the errors tolerance. The Reed Solomon Code [27], denoted $rs(n, k)$, computes corrective information for messages length of 2^n and can correct at least $\frac{n-k}{2}$ errors. The challenge is to find k in order to optimize the overall results. As this kind of code corrects errors on bits, we use the Hamming distance between the users to determine the threshold k .

4.5 Experiments

A signature database of 20 persons, each providing 20 signatures was built. The system is evaluated as follows. The database is divided into two subsets : one (Ref_{base}) for the enrolment containing R randomly selected reference signature per user and one ($Test_{base}$) for the testing (the other $20 - R$ signatures left). The enrolment step is then made. After that, each signature from ($Test_{base}$) is compared to every signals coming from the (Ref_{base}) using the weighted-euclidian distance. Finally, the decision step is applied. A run of 1000 iteration has been done in order to create a large number of different (Ref_{base}). The EER (Equal Error Rate, [25]) computed takes into account all the users in the database.

Results are given in table 1. It shows the EER (in %) for each decision criteria when the enrolment consider 2 to 5 signature per user. Some other experiments showed a minor influence on the ratio H_{max}/H_{min} : between 1.5 to 5, the EER stays around 12% in the DC3 decision mode.

The obtained results are encouraging considering the low complexity of the decision criteria. We need to improve the comprehension of the signal's structure

Table 1. EER (in %) of the three decision criteria for R varying from 2 to 5

R	2	3	4	5
EER (DC1)	20	14	13	13
EER (DC2)	12	13	15	27
EER (DC3)	13	12	15	22

in order to build one (or more) stronger classifier(s) and combine their outputs in the final step. The authentication of some users has proven to be rather difficult (EER around 40%). We think this is related to the time *vs.* frequency problem. Our actual method deals with poor frequency resolution, and the features extracted are not precise enough to efficiently distinguish the users. We need, in the future, to improve our algorithm to solve this problem and to extract new features like the reaction time of the skin.

Compared to other common biometrics, (see table 2 - [22]), our method nears face-based biometric systems.

Table 2. Comparison with the others biometry

Biometric	Keystroke	Finger	Iris	Voice	Tactim
EER (%)	6	< 0.01	< 0.001	3	12

We also ran a benchmark to compare our key-generating method with other biometrics. Table 3 presents this comparison (from [16]). The biometrics are sorted by increasing score. In terms of errors (FAR and FRR), Tactim is better than Fingerprint and behind Iris, but the key’s length we can generate is poorer than the generation based on the Voice. Nevertheless, as we use this key as a PIN, the length is less important than the error rates.

Table 3. Comparison of Key generating methods with the others biometrics

Biometrics	Key stroke	Voice	Fingerprint	Tactim	Iris
KeyLength	12	46	69	24	140
FRR	48,4	20	30	17,4	0,47
FAR	-	-	-	15,2	0

5 The Use of Our Architecture

Any e-transaction is represented by a service, handled by a Service Oriented Architecture (definition and explanation of a SOA can be found in [20]). The kernels of such systems are made to be modular, and by addition of specific components can offer secure services. Almost any SOA has common components (eg: use of SSL protocol) to entrust the communications. But there is a lack of harmonization concerning biometrics components. Many companies propose to incorporate in their system any kind of biometric devices, and this, in many different ways. The construction of a platform-independent biometric component can fulfill this need, and so, increase the usability and the security of the entire existing frameworks. Considering that our component will be used by the SOA to initialize secure communications, we need to focus on the applicative sectors of biometry in general, and of the Tactim concept in particular.

Biometric components are a subset of authentication systems. Their usefulness relies on the permissibility of the security policy of the system. If the services providers do not allow you to share on your own your private access, biometric

components can not be applied. On the opposite, for a voting system, this property is an advantage since the services provider has to start the procedure to give access to someone else.

Regarding our subject, we can apply it only in more critical sectors (we have to assume the environment of the user: at work) since the user has to be in a particular state of mind.

5.1 Component Design and Its Finality

SOA implies modularity and good abstraction of the primary layers (communications, interaction between users and services but also among components). We modify the dGov platform ([28]) to be compliant with the JRMI technology to allow a control of our service through a simple call to an object (abstraction of the mutual authentication). The base (for now, a computer) represents the physical and logical location of this object. It computes, with the help of the biometric kernel, the PIN of the inserted smartcard and can send the information of the service needed by the user to the client.

5.2 Analysis of Possible Attacks

Assuming that a thief steals the mobile agent, he can directly access the public area where the biometric data are stored. Nevertheless, as we store the corrective information, these data are useless : the quantity of this information is inferior to the correction capacity.

Secondly the attack can be lead against the PIN. As nobody knows it, the only possibility is to try randomly, but the smart card prevents attackers from exhaustive attacks (only three attempts).

A more complicated scheme consists in enforcing the user to use the system, but the Tactim biometry is based on specific reaction which can not be established under constraints.

Nevertheless our architecture is sensitive to replay attacks; if someone steals correct features, he can replay them on the sensor's surface and gain the access.

One possibility in order to record features is to corrupt the base. The mutual authentication between the service provider and the base (use of SSL protocol) prevents the process from using a fake base, complicating the task of corruption.

6 Conclusion

This paper describes the promising results of a new kind of biometry. The Tactim project relies on a physiology study which shows that quantifying both the user and his state of mind is feasible. This property, not yet exploited in the different kind of biometry, represents a real opportunity. We have presented each domain linked to a biometric problem, theirs challenges and interactions.

The physiology study has helped to understand how to stimulate the skin reaction and so, has permitted the construction of a sensor prototype. The interpretation of the acquired data represented the more critical point of the overall

system. Signal processing tools and acoustics analysis have shown the region of interest in the signatures. This improvement has led us to develop a features extraction algorithm with stable results (the inner class variance tends to be stable among the users). Nevertheless we have yet to improve the precision of our method; Better resolution in frequencies or new features (establishment time of the reaction) must be an interesting way.

Successfully marrying security and biometry is not an easy task. A classical approach of the biometric kernel (give a "yes" or "no" response) is not acceptable, otherwise the data storage has to be non-invasive. In order to overcome these issues we developed a decision layer which outputs the PIN of the owner's smartcard. The results show that our features are more pertinent than fingerprints for two reasons: we obtain a lower error rate effective given the user's agreement.

On the system design side, we have chosen a hardware architecture built around a personal device (embedded the acquisition and data part). In this scheme, the logical tasks (dealing with: the SOA, the biometric kernel) are delegated to the base which represents the location of the authentication component. There is still a lack of harmonization in biometric components: there is no a global protocol yet. We now need to improve the software protocol in order to offer an abstract layer to the other services.

References

1. Rila, L., Mitchell, C.J.: Security analysis of smartcard to card reader communications for biometric cardholder authentication, in Proc. 5th Smart Card Research and Advanced Application Conference (CARDIS '02), USENIX Association, San Jose, California, (2002) 19–28.
2. Bovenlander, E., van Renesse, R.L.: Smartcards and Biometrics: An Overview. in Computer Fraud and Security Bulletin (1995) 8–12.
3. Rila, L.: Denial of Access in Biometrics-Based Authentication Systems. in Infrastructure Security: International Conference (InfraSec 2002) Bristol, UK, October 1-3 (2002), 19–29.
4. Li, Y-P.: Biometric technology overview. in Nuclear Science and Techniques, **17:2**, (2006) 97–105.
5. Barral C., Coron, J-S. and Naccache, D. : Externalized Fingerprint Matching. ICBA, (2004) 309–215.
6. Hachez, G., Koeune, F. and Quisquater J.: Biometrics, Access Control, Smart Cards: A Not So Simple Combination. in Proc. fourth working conference on smart card research and advanced applications on Smart card research and advanced applications, (2001) 273–288.
7. Nilsson, J., Harris, M.: Match-on-Card for Java Cards. White paper, Precise Biometrics (2004).
8. Teoh Beng Jin, A., Ngo Chek Ling, D. and Alwyn Goh, A.: Biohashing: two factor authentication featuring fingerprint data and tokenised random number, Pattern Recognition, **37:11** (2004) 2245–2255.
9. Cheung, K.H. and Kong, A. and Zhang, D. and Kamel, M. and You, J.,: Revealing the Secret of FaceHashing, ICB06 (2006) 106–112.

10. Kong, A., Cheung, K-H., Zhang, D., Kamel, M. and You, J.: An analysis of Bio-Hashing and its variants, *Pattern Recognition*, **39:7** (2006) 1359–1368.
11. Bolle, R.M., Connell J.H., and Ratha, N.K.: Biometric perils and patches, *Pattern Recognition*, **35:12** (2002) 2727–2738.
12. Peyravian, M., Matyas, S.M., Roginsky, A., and Zunic, N.: Generating user-based cryptographic keys and random numbers, *Computers & Security*, **18:7** (1999) 619–626.
13. Monrose, F., Reiter, M.K. and Wetzel, S.: Password Hardening based on Keystroke Dynamics, *International Journal of Information Security*, **1:1** (2001) 69–83.
14. Monrose, F., Reiter, M.K., Li, Q. and Wetzel, S.: Using voice to generate cryptographic keys. In *Proc. of Odyssey 2001, The speaker recognition workshop* (2001) 237–242
15. Sweeney, P.: *Error Control Coding - From theory to practice*. Wiley (Eds), (2002) ISBN: 047084356X.
16. Hao, F., Anderson, R. and Daugman, J.: Combining cryptography with biometrics effectively. Technical Report UCAM-CL-TR-640 2005.
17. Dodis, Y., Reyzin, L. and Smith A.: Fuzzy extractors : how to generate strong keys from biometrics and other noisy data. *Eurocrypt, Lecture Notes in Computer Science* **3027** (2004) 523–540.
18. Tuyls, P., Goseling, J.: Capacity and Examples of Template-Protecting. *BioAW 2004*, (2004) 158–170.
19. B.J Teoh, A., C.L. Ngo, D. and Goh, A.: Personalised cryptographic key generation based on FaceHashing. *Computers & Security*, **23:7** (2004) 606–614.
20. Endrei, M., Ang, J., Arsanjani, A., Sook, C., Compte, P., Luo M and Newling, T.: *Patterns: Service-Oriented Architecture and Web Services*, IBM redbook (Eds), (2004) ISBN:073845317X
21. Malayath, N., Hermansky, H.: Data-driven spectral basis functions for automatic speech recognition. *Speech Communication* **40:4** (2003) 449–466.
22. Booth, I., Barlow, M. and Warson, B.: Enhancements to DTW and VQ decision algorithms for speaker recognition. *Speech Communication* **13:4** (1993) 427–433.
23. McAulay, R., Quatieri, T. : *Speech Analysis Synthesis Based on a Sinusoidal Representation*. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **34** (1986) 744–754.
24. Brouard, T.: Contribution of dermo-science to secure personal electronic identification. 5th world conference and exhibition on the practical application of biometrics (Biometrics 2002).
25. Daugman, J.: Biometric decision landscapes. Technical Report UCAM-CL-TR-482 (2000).
26. Jain, A., Pankanti, S., Prabhakar, S., Hong, L., Ross, A.: Biometrics: A Grand Challenge. *ICPR* , **2** (2004) 935–942.
27. Bruck, J., Naor M.: The hardness of decoding linear codes with preprocessing. *IEEE Trans. Inform. Theory* (1990) 381-385.
28. Sérgio Nunes, DGOV - A secure and generic services platform for the e-Government. Master Thesis. In preparation.
29. Cauchie, S., Brouard, T., De sousa, S.: Deux approches de l'acquisition et du traitement d'un signal biométrique cutané, *Proc. of Capteurs 2005*, (2005) 101–111.

Improving the Dynamic ID-Based Remote Mutual Authentication Scheme

Eun-Jun Yoon and Kee-Young Yoo*

Department of Computer Engineering, Kyungpook National University,
1370 Sankyuk-Dong, Buk-Gu, Daegu 702-701, South Korea
Tel.: +82-53-950-5553; Fax: +82-53-957-4846
ejyoon@infosec.knu.ac.kr, yook@knu.ac.kr

Abstract. In 2005, Liao et al. pointed out the weaknesses of Das et al.'s dynamic ID-based remote user authentication scheme using smart cards, and then proposed a slight modification thereof to overcome these weaknesses. The current paper, however, demonstrates that Liao et al.'s scheme is still vulnerable to reflection attacks, privileged insider's attacks, and impersonation attacks by using lost or stolen smart card. Then, we present an improvement to the scheme in order to isolate such problems.

Keywords: Authentication, Password, Dynamic ID, Smart card, Reflection attack.

1 Introduction

Password authentication is a simple and convenient authentication mechanism that allows a legal user to login to remote systems. A number of researchers [1][2][3][4][5] have proposed password authentication schemes for secure login of legal users. However, all these schemes are based on static login ID, which is vulnerable to leaking partial information about a user's login message to an adversary. One solution to ID-theft is employing a dynamic ID for each login.

In 2004, Das et al. [6] proposed a dynamic ID-based remote user authentication scheme using smart cards. The scheme has the following advantages: (1) It allows the users to choose and change their passwords freely. (2) It does not maintain any verifier tables in the remote system. (3) The remote user authentication scheme is secure against ID-theft, replay attacks, forgery attacks, insider attacks, and stolen-verifier attacks. In 2005, Liao et al. [7], however, showed that Das et al.'s scheme has three security weaknesses as follows: (1) It cannot protect against guessing attacks. (2) It cannot achieve mutual authentication. (3) Passwords can be revealed by remote systems. Liao et al. proposed a slight modification of Das et al.'s scheme. They claimed that their proposed scheme not only achieves Das et al.'s advantages; it also enhances Das et al.'s security by removing the security weaknesses.

* Corresponding author.

Unlike Liao et al.'s claims, their scheme is still vulnerable to a reflection attack [8] and a privileged insider's attack [9]. Therefore, the current paper demonstrates that Liao et al.'s scheme is still vulnerable to reflection attacks, privileged insider's attacks, and impersonation attacks using lost or stolen smart cards. Then, we present an improvement to the scheme in order to remove such problems.

This paper is organized as follows: Section 2 briefly reviews Liao et al.'s dynamic ID-based remote user authentication scheme using smart cards; then Section 3 discusses its weaknesses. Our proposed scheme is presented in Section 4, while Section 5 discusses the security and efficiency of the proposed scheme. Our conclusions are presented in Section 6.

2 Review of Liao et al.'s Scheme

This section briefly reviews Liao et al.'s dynamic ID-based remote mutual authentication scheme using smart cards [7]. Some of the notations used in this paper are defined as follows:

- U : The user
- PW : The password of U
- S : The remote system
- x : The secret key of S
- y : The secret number of S stored in each user's smart card
- T : A time-stamp
- $h(\cdot)$: A one-way hash function
- \oplus : Bit-wise XOR operation
- \parallel : Concatenation

Liao et al.'s scheme consists of two phases: a registration phase and an authentication phase. Figure 1 shows Liao et al.'s authentication scheme. The scheme works as follows:

2.1 Registration Phase

When U requests to register with S , S performs this phase only once as follows:

1. U freely chooses a password PW and computes $h(PW)$. He/she submits his/her identity ID and $h(PW)$ to S through a secure channel.
2. S then computes $N = h(PW) \oplus h(x \parallel ID)$.
3. S stores $(N, y, h(\cdot))$ into a smart card and then sends the smart card to U through a secure channel.

2.2 Authentication Phase

When U wants to login S , S authenticates U as follows:

1. U inserts his/her smart card into the card reader of a terminal, and keys in his/her PW . Then, the smart card can compute a dynamic ID as $CID = h(PW) \oplus h(N \oplus y \oplus T)$, $B = h(CID \oplus h(PW))$, and $C = h(T \oplus N \oplus B \oplus y)$, where T is a time-stamp.

2. U sends (CID, N, C, T) to S .
3. Upon receiving the login request at the time T' , S verifies whether $(T' - T) \leq \Delta T$. If it holds, S accepts the login request of U , where ΔT is an expected valid time interval. Then, S computes $h(PW) = CID \oplus h(N \oplus y \oplus T)$ and $B = h(CID \oplus h(PW))$, and checks if $C = h(T \oplus N \oplus B \oplus y)$. If it holds, S accepts U to login to the system. Otherwise, S rejects it. Then S computes $D = h(T^* \oplus N \oplus B \oplus y)$, where T^* is a time-stamp.
4. S sends (D, T^*) to U .
5. Upon receiving a reply message at the time T'' , U verifies whether $(T'' - T^*) \leq \Delta T$, where ΔT is an expected valid time interval. If it holds, U computes $h(T^* \oplus N \oplus B \oplus y)$ and compares it with the received D . If it holds, U can be sure he or she is communicating with the actual S .

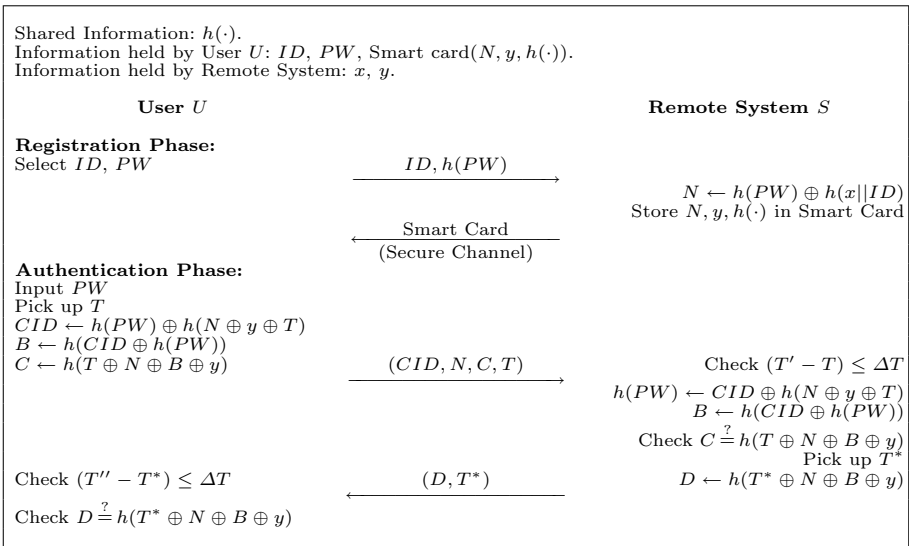


Fig. 1. Liao et al.'s Authentication Scheme

3 Cryptanalysis of Liao et al.'s Scheme

This section shows that Liao et al.'s authentication scheme is vulnerable to reflection attacks, insider attacks, and impersonation attacks using lost or stolen smart cards.

3.1 Reflection Attack

A reflection attack [8] is a potential way of attacking a challenge-response authentication system which uses the same protocol in both directions. The basic idea is to trick the target into providing the answer to its own challenge. Consider a scenario of a reflection attack in Liao et al.'s scheme. In the authentication

phase, if attacker A has intercepted and blocked a message transmitting in Step (2), i.e., (CID, N, C, T) , he or she can impersonate the remote system and send (D, T^*) to U in Step (4) of the authentication phase, where $D = C$ and $T^* = T$ is the current timestamp. Upon receiving the second item of the received message, i.e., T^* , U will believe T^* is a valid timestamp because $(T'' - T^*) \leq \Delta T$. Then U will compute $h(T^* \oplus N \oplus B \oplus y)$. Note that Step (3) of the authentication phase is skipped by attacker A . Since the computed result equals the first item of the received message, i.e., $D = C = h(T^* \oplus N \oplus B \oplus y)$, where $T^* = T$, U is fooled into believing that the attacker is the legal remote system. Since U cannot actually authenticate the remote system's identity, Liao et al.'s authentication scheme fails to provide mutual authentication. Fig. 2 depicts the message transmission of the reflection attack.

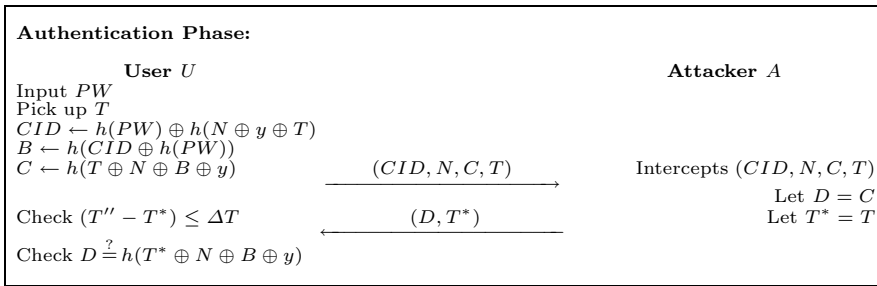


Fig. 2. Reflection attack on Liao et al.'s scheme

3.2 Insider Attack

Liao et al. claimed that U 's password PW cannot be revealed by remote system S in the registration phase because it is protected by a one-way hash function. Assuming that user U picks a predictable password, PW , then unlike their claim, in the registration phase of Liao et al.'s scheme, U 's password PW , however, will be revealed to the remote system because its hashed value $h(PW)$ is transmitted to the system. That is, the system first guesses password PW^* and checks if $h(PW) = h(PW^*)$ by using an off-line password guessing attack. If it holds, S can get U 's password PW . Otherwise, S performs it until $h(PW) = h(PW^*)$. In practice, users offer the same password to access several servers for their convenience. Thus, the privileged insider of the remote system may try to use PW to impersonate U to login to the other systems that U has registered with outside this system [9]. If the targeted outside system adopts the normal password authentication scheme, it is possible that the privileged insider of the system can successfully impersonate U to login to it by using PW . Although it is also possible that all the privileged insiders of the system are trusted and U does not use the same password to access several systems, the implementers and the users of the scheme should be aware of such a potential weakness.

3.3 Impersonation Attack by Using Lost or Stolen Smart Card

Liao et al.'s scheme is vulnerable to impersonation attacks using a lost or stolen smart cards. Namely, a user can get authenticated to a remote system even if s/he doesn't have the valid password. Precisely, if an attacker gets a user's smart card, he can input any string str in the smart card as the user's password, and can still get authenticated. Needless to say this is a serious problem. Fig. 3 depicts the message transmission of an impersonation attack using a lost or stolen smart card.

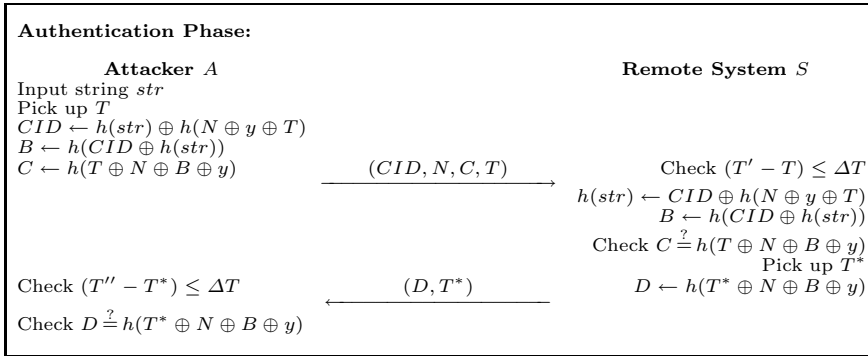


Fig. 3. Impersonation attack by using lost or stolen smart card on Liao et al.'s scheme

4 Countermeasures

This section proposes an enhancement to Liao et al.'s scheme that eliminates the security flaws described in the previous section. Our improved scheme not only possesses the advantages of their scheme, it also enhances the security of their scheme. Our scheme is also divided into the two phases of registration and authentication. Fig. 4 illustrates the proposed remote user authentication scheme. To resist such attacks, the proposed phases perform as follows:

4.1 Registration Phase

When a new user U wants to register with the remote system S , he/she performs this phase only once. S will issue a smart card to U after this phase is done. The steps are as follows:

1. U freely chooses a password PW and computes $h(PW||R)$, where R is randomly chosen nonce by U . He/she submits his/her identity ID and $h(PW||R)$ to S through a secure channel.
2. S then computes $N = h(PW||R) \oplus h(x||ID)$ and $K = h(PW||R) \oplus h(N||y)$.
3. S stores $(N, y, K, h(\cdot))$ into a smart card and then sends the smart card to U through a secure channel.
4. U enters R into his/her smart card.

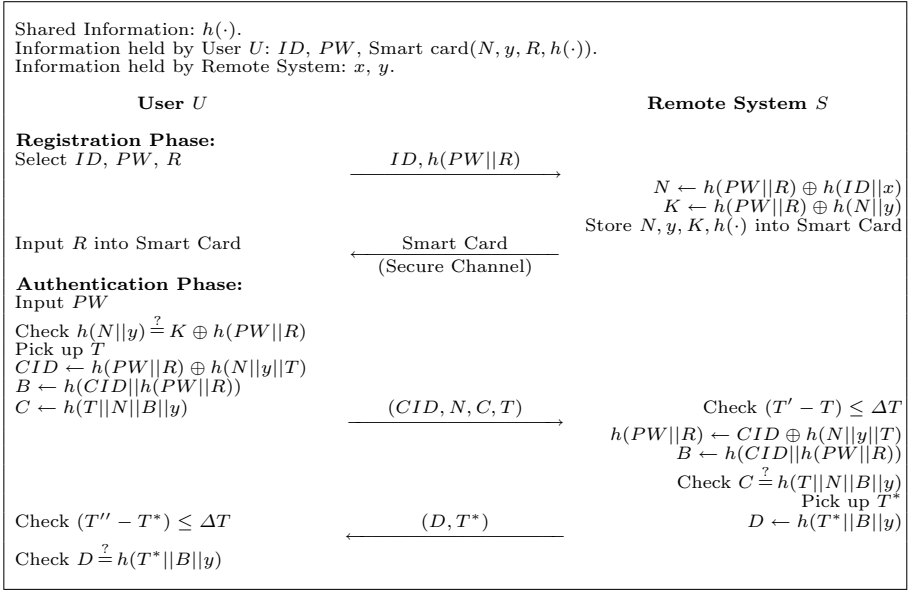


Fig. 4. Proposed Authentication Scheme

4.2 Authentication Phase

In this phase, when U wants to login S , S can authenticate U . The steps of this phase are as follows:

1. U inserts his/her smart card into the card reader of a terminal, and keys in his/her PW . Then, the smart card computes $h(PW||R)$ and extracts $h(N||y)$ by computing $K \oplus h(PW||R)$. The smart card computes $h(N||y)$ by using stored N and y , and compares it with extracted hash value $h(N||y)$. If it is equal, the smart card computes a dynamic ID as $CID = h(PW||R) \oplus h(N||y||T)$, $B = h(CID||h(PW||R))$, and $C = h(T||N||B||y)$, where T is a time-stamp.
2. U sends (CID, N, C, T) to S .
3. Upon receiving the login request at the time T' , S verifies if whether $(T' - T) \leq \Delta T$. If it holds, S accepts the login request of U , where ΔT is an expected valid time interval. Then, S computes $h(PW||R) = CID \oplus h(N||y||T)$, $B = h(CID||h(PW||R))$, and checks if $C = h(T||N||B||y)$. If it holds, S allows U to login to the system. Otherwise, S rejects it. Then S computes $D = h(T^*||B||y)$, where T^* is a time-stamp.
4. S sends (D, T^*) to U .
5. Upon receiving the reply message at the time T'' , U verifies whether $(T'' - T^*) \leq \Delta T$, where ΔT is an expected valid time interval. If it holds, U computes $h(T^*||B||y)$ and compares it with the received D . If it holds, U can be sure that s/he is communicating with the actual S .

5 Security and Efficiency Analysis

This section discusses the security and efficiency features of the proposed authentication scheme.

5.1 Security Analysis

This subsection provides the proof of correctness of the proposed authentication scheme. First, the security terms [8] needed for the analysis of the proposed scheme are defined as follows:

Definition 1. *A weak password (PW) has a value of low entropy, which can be guessed in polynomial time.*

Definition 2. *A strong secret key (x) has a value of high entropy, which cannot be guessed in polynomial time.*

Definition 3. *A secure one-way hash function $y = h(x)$ is where given x to compute y is easy and given y to compute x is hard.*

Given the above definitions, the following analyzes the security of the proposed authentication scheme:

1. *The proposed scheme prevents the reflection attack in Liao et al.'s scheme:* The reflection attack on Liao et al.'s scheme can succeed due to the symmetric structure of the messages (e.g. $C = h(T \oplus N \oplus B \oplus y)$ and $D = h(T^* \oplus N \oplus B \oplus y)$) exchanged between the user U and the remote system. However, the proposed scheme can prevent reflection attacks like those in Liao et al.'s scheme because of the different message structure between $C = h(T||N||B||y)$ and $D = h(T^*||B||y)$. Thus, the proposed scheme prevents reflection attacks such as in Liao et al.'s scheme.
2. *The proposed scheme can resist an insider attack:* Since U registers to the server by presenting $h(PW||R)$ instead of $h(PW)$, the insider of the server cannot obtain PW without knowing random nonce R using an off-line password guessing attack.
3. *The proposed scheme can resist an impersonation attack using a lost or stolen smart card:* Suppose legal users lost their smart card or an attacker steals the smart card for a short duration and makes a duplicate of it. If an attacker inputs any string PW' in the smart card as the user's password, the attack cannot pass the smart card verification process, step 1 of the authentication phase, because the attacker does not know the legal user's password PW and the smart card checks $h(N||y) = K \oplus h(PW'||R)$. Thus, the proposed scheme prevents an impersonation attack using lost or stolen smart cards as Liao et al.'s scheme is vulnerable to.
4. *The proposed scheme can resist replay attacks:* For replay attacks, neither the replay of an old login message (CID, N, C, T) in the authentication phase will work, as it will fail in Steps 3 due to the time interval $(T' - T) \leq \Delta T$.

5. *The proposed scheme can resist password guessing attacks:* In the registration phase, S embedded each user's identity ID in N . Assume that a user wants to use password guessing attacks in the proposed scheme. It cannot attack the proposed scheme because N includes each user's ID and random number R .
6. *The proposed scheme can achieve mutual authentication:* In the authentication phase, S can authenticate U . U can authenticate S in Step 5 of authentication phase because only valid S can compute $h(T^*||B||y)$. Therefore, the proposed scheme can achieve mutual authentication.

5.2 Efficiency Analysis

Comparisons between Dae et al's scheme [6], Liao et al's scheme [7], and our proposed scheme are shown in Table 1. To analyze the computational complexity of the proposed scheme, we define the notation T_h , which is the time for computing one-way hash function.

In Das et al.'s scheme, S computes N that requires $2 \times T_h$ in the registration phase. In the authentication phase, U computes CID that requires $2 \times T_h$, computes B that requires $1 \times T_h$, and computes C that requires $1 \times T_h$. In the same phase, S computes $CID \oplus h(N \oplus y \oplus T)$ that requires $1 \times T_h$, computes B that requires $1 \times T_h$, and computes $h(T \oplus N \oplus B \oplus y)$ that requires $1 \times T_h$.

In Liao et al's scheme, U computes $h(PW)$ that requires $1 \times Th$ and S computes N that requires $1 \times T_h$ in registration phase. In authentication phase of Liao et al's scheme, U computes CID that requires $2 \times T_h$, computes B requires $1 \times T_h$, and computes C that requires $1 \times T_h$. U verifies D that requires $1 \times Th$. In the same phase, S computes $CID \oplus h(N \oplus y \oplus T)$ that requires $1 \times T_h$, computes B requires $1 \times T_h$, computes $h(T \oplus N \oplus B \oplus y)$ that requires $1 \times T_h$, and computes $h(T^* \oplus N \oplus B \oplus y)$ requires $1 \times T_h$.

In the proposed scheme, U computes $h(PW||R)$ that requires $1 \times Th$ and S computes N and K that require $2 \times T_h$ in registration phase. In authentication phase of proposed scheme, U computes $h(N||y)$ that requires $1 \times T_h$, $h(PW||R)$ requires $1 \times T_h$, CID requires $1 \times T_h$, computes B that requires $1 \times T_h$, and computes C that requires $1 \times T_h$. U verifies D requires $1 \times Th$. In the same phase, S computes $h(N||y||T)$ that requires $1 \times T_h$, B requires $1 \times T_h$, computes $h(T||N||B||y)$ that requires $1 \times T_h$, and computes $h(T^*||B||y)$ that requires $1 \times T_h$.

We can see that the number of hash operation is increased by only two in our scheme compared with Liao et al.'s scheme. It does not add many additional

Table 1. A comparison of computation costs

Computational type	Das et al.'s Scheme [6]	Liao et al.'s Scheme [7]	Proposed Scheme
Registration Phase	$2 \times T_h$	$2 \times T_h$	$3 \times T_h$
Authentication Phase	$7 \times T_h$	$9 \times T_h$	$10 \times T_h$

computational costs. Therefore, the proposed scheme is not only simple but also enhances security.

6 Conclusions

The current paper demonstrated that Liao et al.'s scheme is vulnerable to reflection attacks, privileged insider's attacks, and impersonation attacks using lost or stolen smart cards, and then presented an improved scheme in order to remove such problems. As a result, in contrast to Liao et al.'s scheme, the proposed scheme is able to provide greater security.

Acknowledgements

This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

References

1. Lamport, L.: Password Authentication with Insecure Communication. *Communications of the ACM*. Vol. 24. (1981) 770-772
2. Hwang, M.S., Lee, C.C., Tang, Y.L.: A Simple Remote User Authentication Scheme. *Mathematical and Computer Modelling*. Vol. 36. (2002) 103-107
3. Lee, C.C., Hwang, M.S., Yang, W.P.: Flexible Remote User Authentication Scheme Using Smart Cards. *ACM Operating Systems Review*. Vol. 36. No. 3. (2002) 46-52
4. Li, L.H., Lin, I.C., Hwang, M.S.: A Remote Password Authentication Scheme for Multiserver Architecture Using Neural Networks. *IEEE Transactions on Neural Networks*. Vol. 12. No. 6. (2001) 1498-1504
5. Shen, J.J., Lin, C.W., Hwang, M.S.: A Modified Remote User Authentication Scheme Using Smart Cards. *IEEE Transactions on Consumer Electronics*. Vol. 49. No. 2. (2003) 414-416
6. Das, M.L., Saxena, A., Gulati, V.P.: A Dynamic ID-based Remote User Authentication Scheme. *IEEE Transactions on Consumer Electronics*. Vol. 50. No. 2. (2004) 629-631
7. Liao, I.E., Lee, C.C., Hwang, M.S.: Security Enhancement for a Dynamic ID-based Remote User Authentication Scheme. *IEEE Computer Society. Proceedings of the International Conference on Next Generation Web Services Practices (NWeSP 2005)*. (2005) 437-440
8. Menezes, A.J., Oorschot, P.C., Vanstone, S.A.: *Handbook of Applied Cryptograph*. CRC Press. New York. (1997)
9. Ku, W.C., Chuang, H.M., Tsaur, M.J.: Vulnerabilities of Wu-Chieu's Improved Password Authentication Scheme Using Smart Cards. *IEICE Trans. Fundamentals*. Vol. E88-A. No. 11. (November 2005) 3241-3243

Security Enhancement of a Remote User Authentication Scheme Using Smart Cards*

Youngsook Lee, Junghyun Nam, and Dongho Won**

Information Security Group, Sungkyunkwan University, Korea
{yslee, jhnam, dhwon}@security.re.kr

Abstract. Designing cryptographic protocols well suited for today's distributed large networks poses great challenges in terms of cost, performance, user convenience, functionality, and above all security. As has been pointed out for many years, even designing a two-party authentication scheme is extremely error-prone. This paper discusses the security of Lee et al.'s remote user authentication scheme making use of smart cards. Lee et al.'s scheme was proposed to solve the security problem with Chien et al.'s authentication scheme and was claimed to provide mutual authentication between the server and the remote user. However, we demonstrate that Lee et al.'s scheme only achieves unilateral authentication — only the server can authenticate the remote user, but not vice versa. In addition, we recommend changes to the scheme that fix the security vulnerability.

Keywords: Authentication scheme, password, smart card, parallel session attack, reflection attack.

1 Introduction

A mutual authentication scheme is a two-party protocol designed to allow the communicating parties to confirm each other's identity over a public, insecure network. Authentication schemes are necessary for secure communication because one needs to know with whom he or she is communicating before sending some sensitive information. Achieving any form of authentication inevitably requires some secret information to be established between the communicating parties in advance of the authentication stage. Cryptographic keys, either secret keys for symmetric cryptography or private/public keys for asymmetric cryptography, may be one form of the underlying secret information pre-established between the parties. However, these high-entropy cryptographic keys are random in appearance and thus are difficult for humans to remember, entailing a significant amount of administrative work and costs. Eventually, it is this drawback that password-based authentication came to be widely used in reality. Passwords

* This work was supported by the Korean Ministry of Information and Communication under the Information Technology Research Center (ITRC) support program supervised by the Institute of Information Technology Assessment (IITA).

** Corresponding author.

are drawn from a relatively small space like a dictionary, and are easier for humans to remember than cryptographic keys with high entropy.

The possibility of password-based user authentication in remotely accessed computer systems was explored as early as the work of Lamport [10]. Due in large part to the practical significance of password-based authentication, this initial work has been followed by a great deal of studies and proposals, including solutions using multi-application smart cards [4,14,8,13,5,16,15]. In a typical password-based authentication scheme using smart cards, remote users are authenticated using their smart card as an identification token; the smart card takes as input a password from a user, recovers a unique identifier from the user-given password, creates a login message using the identifier, and then sends the login message to the server, who then checks the validity of the login request before allowing access to any services or resources. This way, the administrative overhead of the server is greatly reduced and the remote user is allowed to remember only his password to log on. Besides just creating and sending login messages, smart cards support mutual authentication where a challenge-response interaction between the card and the server takes place to verify each other's identity. Mutual authentication is a critical requirement in most real-world applications where one's private information should not be released to anyone until mutual confidence is established. Indeed, phishing attacks [1] are closely related to the deficiency of server authentication, and are a growing problem for many organizations and Internet users.

The experience has shown that the design of secure authentication schemes is not an easy task to do, especially in the presence of an active intruder; there is a long history of schemes for this domain being proposed and subsequently broken by some attacks (e.g., [6,2,3,12,7,16,15,9]). Therefore, authentication schemes must be subjected to the strictest scrutiny possible before they can be deployed into an untrusted, open network. In 2000, Sun [13] proposed a remote user authentication scheme using smart cards. Compared with the earlier work of Hwang and Li [8], this scheme is extremely efficient in terms of the computational cost since the protocol participants perform only a few hash function operations. In 2002, Chien et al. [5] presented another remote user authentication scheme which improves on Sun's scheme in two ways; it provides mutual authentication and allows users to freely choose their passwords. However, Hsu [7] has pointed out that Chien et al.'s scheme is vulnerable to a parallel session attack; an intruder can masquerade as a legitimate user by using server's response for an honest session as a valid login message for a fake, parallel session. To patch this security vulnerability, Lee et al. [11] have recently presented a slightly modified version of Chien et al.'s scheme, and have claimed, among others, that their modified version achieves mutual authentication between the server and the remote user. But, unlike the claim, their modification only achieves unilateral authentication; only the server can authenticate the remote user, but not vice versa. In this paper, we demonstrate this by showing that Lee et al.'s revised scheme is still insecure against a reflection attack. Besides reporting the reflection attack on

Lee et al.'s scheme, we also figure out what has gone wrong with the scheme and how to fix it.

The remainder of this paper is organized as follows. We begin by reviewing Chien et al.'s remote user authentication scheme and its weakness in Section 2. We continue in Section 3 with a description of Lee et al.'s scheme. Then, we present a reflection attack on Lee et al.'s scheme in Section 4, and show how to prevent the attack in Section 5. Finally, we conclude this work in Section 6.

2 Review of Chien et al.'s Authentication Scheme and Its Weakness

This section reviews Chien et al.'s remote user authentication scheme [5] and Hsu's parallel session attack [7] on it. Chien et al.'s scheme consists of three phases: the registration phase, the login phase, and the verification phase. The registration phase is performed only once per user when a new user registers itself with the server. The login and authentication phases are carried out whenever a user wants to gain access to the server. A pictorial view of the scheme at a high level of abstraction is given in Fig. 1, where a dashed line indicates a secure channel, and a more detailed description follows.

2.1 Chien et al.'s Authentication Scheme

Registration Phase. Let x be the secret key of the authentication server (AS), and h be a secure one-way hash function. A user U_i , who wants to register with the server AS , chooses its password PW_i at will and submits a registration request, consisting of its identity ID_i and password PW_i , to the server AS via a secure channel. Then AS computes

$$X_i = h(ID_i \oplus x) \quad \text{and} \quad R_i = X_i \oplus PW_i$$

and issues a smart card containing $\langle R_i, h^* \rangle$ to U_i , where h^* denotes the description of the hash function h .

Login Phase. When U_i wants to log in to the system, he inserts his smart card into a card reader and enters his identity ID_i and password PW_i . Given ID_i and PW_i , the smart card computes

$$X_i = R_i \oplus PW_i \quad \text{and} \quad C_1 = h(X_i \oplus T_1)$$

where T_1 is the current timestamp. The smart card then sends the login request message $\langle ID_i, T_1, C_1 \rangle$ to the server AS .

Verification Phase. With the login request message $\langle ID_i, T_1, C_1 \rangle$, the scheme enters the verification phase during which AS and U_i perform the following steps:

Step 1. Upon receiving the message $\langle ID_i, T_1, C_1 \rangle$, the server AS checks that:
 (1) ID_i is valid, (2) $T_2 - T_1 \leq \Delta T$, where T_2 is the timestamp when AS

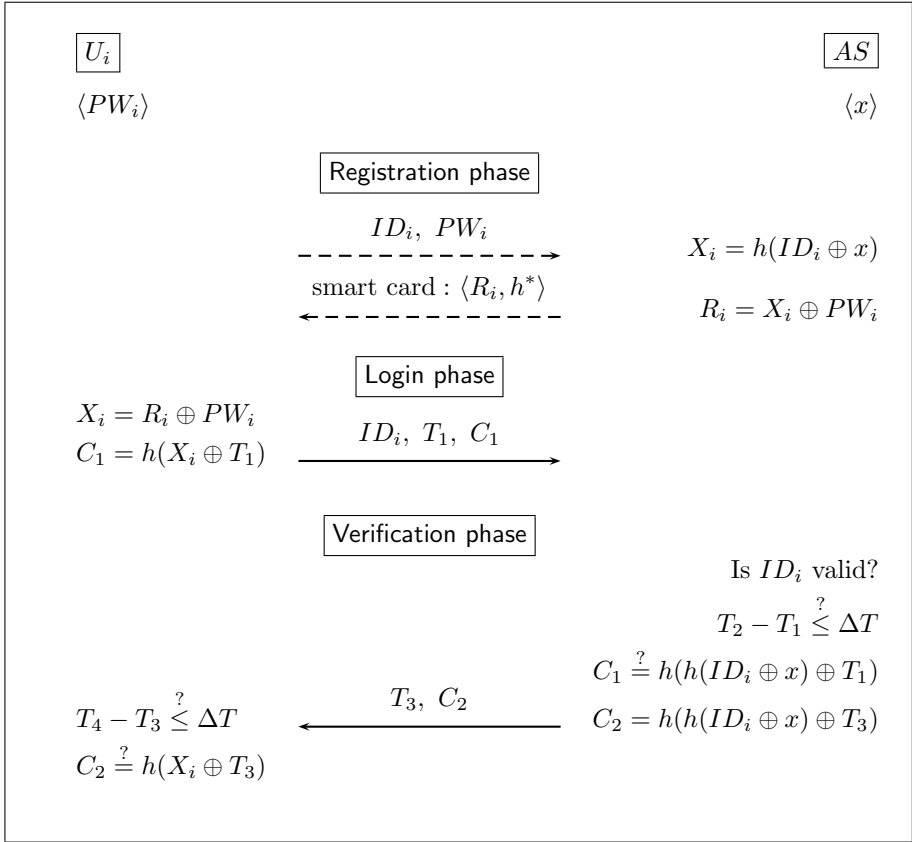


Fig. 1. Chien et al.'s remote user authentication scheme

received the login request message and ΔT is the maximum allowed time difference between T_1 and T_2 , and, finally, (3) C_1 is equal to $h(h(ID_i \oplus x) \oplus T_1)$. If any of these are untrue, AS rejects the login request and aborts the protocol. Otherwise, AS accepts the login request.

Step 2. Now, AS obtains the current timestamp T_3 , computes $C_2 = h(h(ID_i \oplus x) \oplus T_3)$, and sends the response message $\langle T_3, C_2 \rangle$ to user U_i .

Step 3. After receiving the message $\langle T_3, C_2 \rangle$ from AS , user U_i checks that: (1) $T_4 - T_3 \leq \Delta T$, where T_4 is the timestamp when U_i received the response message $\langle T_3, C_2 \rangle$, and (2) C_2 equals $h(X_i \oplus T_3)$. If both of these conditions hold, U_i believes that he is talking to the authentic server. Otherwise, U_i aborts his login attempt.

2.2 Hsu's Attack on Chien et al.'s Scheme

As already mentioned, Hsu [7] showed that Chien et al.'s remote user authentication scheme is vulnerable to a parallel session attack through which an intruder

E is easily able to gain access to the server by disguising herself into a legitimate user U_i . In Hsu’s attack, E simply eavesdrops on AS ’s response message $\langle T_3, C_2 \rangle$ for an honest session between U_i and AS , and immediately starts a parallel session sending the forged login request message $\langle ID_i, T_3, C_2 \rangle$ to the server AS . Since C_2 equals $h(h(ID_i \oplus x) \oplus T_3)$, AS believes that the login request message $\langle ID_i, T_3, C_2 \rangle$ comes from another instance of U_i as long as the message arrives at AS before the timer expires.

The vulnerability of Chien et al.’s scheme to this parallel session attack is mainly because that two authenticators C_1 and C_2 exchanged between two authenticating parties are computed using the same cryptographic expression: $h(h(ID_i \oplus x) \oplus timestamp)$. Indeed, this is a well-known fundamental flaw of authentication schemes that allows an intruder to use messages going to one direction to construct forged — but still valid — messages going to the opposite direction [6,2].

3 Lee et al.’s Authentication Scheme

To thwart the parallel session attack, Lee et al. [11] have recently presented an improved version of Chien et al.’s scheme. The registration and login phases of Lee et al.’s scheme are the same as those of Chien et al.’s scheme. Furthermore, the only difference between the verification phases of two schemes is in the

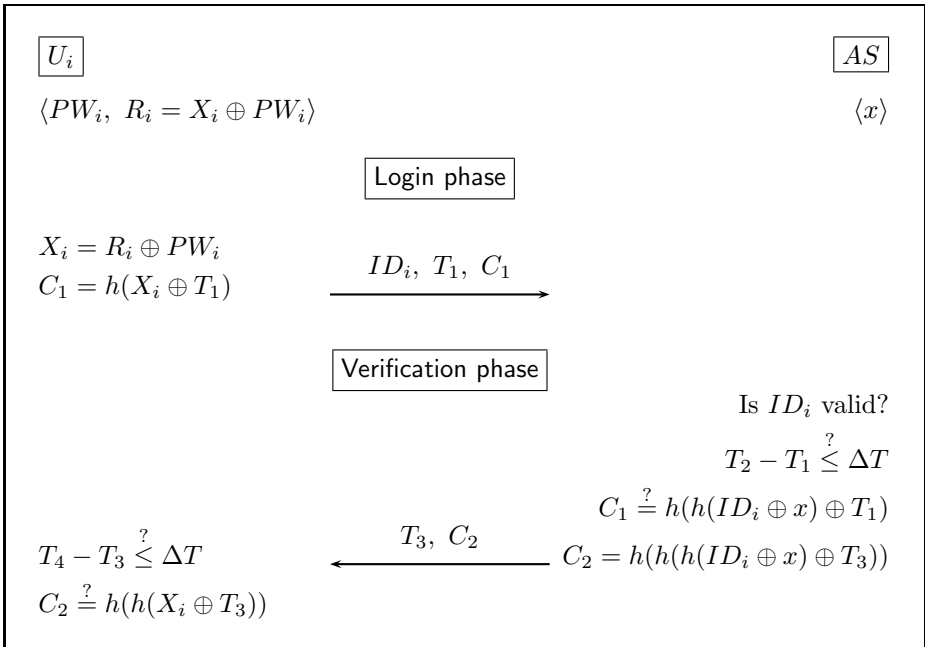


Fig. 2. Lee et al.’s remote user authentication scheme

computation of C_2 , i.e., $C_2 = h(h(ID_i \oplus x) \oplus T_3)$ versus $C_2 = h(h(h(ID_i \oplus x) \oplus T_3))$). A high level depiction of the verification phase of Lee et al.'s scheme is given in Fig. 2 and a more detailed description follows.

Verification Phase. The following steps are performed with the login request message $\langle ID_i, T_1, C_1 \rangle$ being sent to AS by U_i :

Step 1. Upon receiving $\langle ID_i, T_1, C_1 \rangle$, the server AS acquires the current timestamp T_2 and verifies that: (1) ID_i is valid, (2) $T_2 - T_1 \leq \Delta T$, where T_1 and ΔT are as defined in Chien et al.'s scheme, and (3) C_1 equals $h(h(ID_i \oplus x) \oplus T_1)$. If all of these conditions hold, AS accepts the login request. Otherwise, AS rejects it and aborts the protocol.

Step 2. AS generates a new timestamp T_3 , computes C_2 as $C_2 = h(h(h(ID_i \oplus x) \oplus T_3))$, and sends the response message $\langle T_3, C_2 \rangle$ to user U_i .

Step 3. Upon receipt of the response $\langle T_3, C_2 \rangle$, user U_i generates a new timestamp T_4 and checks that: (1) $T_4 - T_3 \leq \Delta T$ and (2) C_2 is equal to $h(h(X_i \oplus T_3))$ where $X_i = h(ID_i \oplus x)$. If both of these conditions hold, U_i believes AS as authentic. Otherwise, U_i aborts his login attempt.

It is straightforward to see that Lee et al.'s authentication scheme is secure against Hsu's parallel session attack since the intruder can no longer use the server's response C_2 in forging a valid login request message unless she can invert the hash function h .

4 Attack on Lee et al.'s Authentication Scheme

Unfortunately, Lee et al.'s remote user authentication scheme provides only unilateral authentication. To show this, we present a reflection attack where an intruder impersonates AS to U_i . The attack scenario is outlined in Fig. 3 and is described in more detail as follows:

1. As usual, the verification phase begins when user U_i sends the login request message $\langle ID_i, T_1, C_1 \rangle$ to the server AS .
2. But, the intruder E intercepts this login request message and computes $C_E = h(C_1)$. E then immediately sends the forged response message $\langle T_1, C_E \rangle$ to user U_i alleging that it comes from the server AS .
3. The timestamp T_1 that U_i receives from E , who is posing as AS , is in fact the timestamp sent out by U_i himself. However, U_i cannot detect this fact since the scheme does not require U_i to check whether or not the timestamp received from the server equals the one sent by U_i himself; to follow the specification of the scheme is all that he can and should do. Hence, everything proceeds as usual; U_i checks that $T_4 - T_1 \leq \Delta T$ and C_E equals $h(h(X_i \oplus T_1))$. Since C_E is equal to $h(h(X_i \oplus T_1))$, the forged response message $\langle T_1, C_E \rangle$ will pass the verification test as long as the condition $T_4 - T_1 \leq \Delta T$ holds, which is indeed the case.

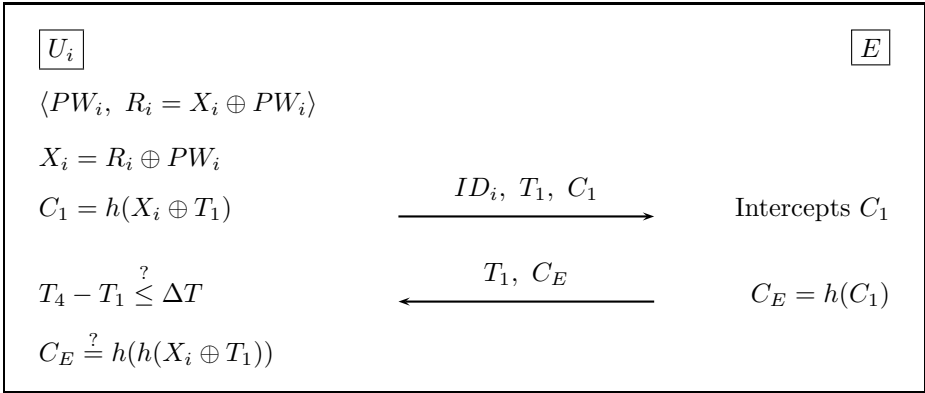


Fig. 3. Attack on Lee et al.’s authentication scheme

The basic idea of our attack is essentially similar to that of Hsu’s parallel session attack on Chien et al.’s scheme — when an honest protocol participant sends a message to his authenticating party, the intruder eavesdrops or intercepts the message and sends it (or a modified version of it) back to the message originator.

Note that a similar attack scenario as above can be also applied to Chien et al.’s scheme. Hence, we can say that the original Chien et al.’s scheme does not guarantee any kind of authentication, either user-to-server authentication or server-to-user authentication. The problem with Lee et al.’s authentication scheme is that it fixes only one of two problems and thus fails to achieve mutual authentication.

5 Preventing the Reflection Attack

We now figure out what is wrong with Lee et al.’s scheme and how to fix it, in the hope that no similar mistakes are made in the future.

5.1 Flaw in the Scheme

Lee et al. [11] claimed that their scheme prevents the intruder from impersonating AS to U_i . In support of this claim, they argue that the intruder cannot compute the server’s response C_2 because she does not know the secret value $X_i = h(ID_i \oplus x)$. But, this claim is flawed. To compute $C_2 = h(h(h(ID_i \oplus x) \oplus timestamp))$, the intruder does not need to know X_i , rather it suffices to know the value $h(h(ID_i \oplus x) \oplus timestamp)$. It is this flaw that has led us to present the reflection attack in which the intruder can easily succeed in impersonating AS to U_i . Using this flaw, the intruder E intercepts login request message C_1 and then immediately sends $\langle T_1, C_E = h(C_1) \rangle$ back to U_i . We emphasize again that C_E is a valid response as long as it arrives within the time window.

5.2 Countermeasure

One obvious solution to this vulnerability is to modify the cryptographic expressions used in computing C_1 and C_2 so that it is infeasible for the intruder to compute C_2 from C_1 . We therefore change the computation of C_1 and C_2 to:

$$C_1 = h(ID_i, h(ID_i \oplus x), T_1)$$

and

$$C_2 = h(ID_i, C_1, h(ID_i \oplus x), T_3).$$

With this modification, it would be impossible for the intruder to mount the reflection attack. The intruder, who wants to impersonate AS to U_i , can no longer forge a valid server's response from the login request message $\langle ID_i, T_1, C_1 \rangle$ because C_2 cannot be computed from C_1 without knowing the secret value $h(ID_i \oplus x)$. Our modification also prevents Hsu's parallel session attack. Even if the intruder eavesdrops on the server's response message $\langle T_3, C_2 \rangle$, she is unable to construct from it a valid login request message because C_1 cannot be computed from C_2 without knowing $h(ID_i \oplus x)$. Therefore, neither Hsu's parallel session attack nor our reflection attack can be applied to the fixed scheme.

6 Conclusion

A password-based scheme for remote user authentication using smart cards was proposed in the recent work of Lee et al. [11]. Despite its many merits, Lee et al.'s scheme only achieves unilateral authentication unlike the claim that the scheme provides mutual authentication. To demonstrate this, we have shown that the scheme is vulnerable to a reflection attack in which an intruder is easily able to impersonate the authentication server to users. In addition, we have recommended a small change to the scheme that can address the identified security problem.

References

1. Anti-Phishing Working Group, <http://www.antiphishing.org>.
2. R. Bird, I. Gopal, A. Herzberg, P. A. Janson, S. Kutten, R. Molva, and M. Yung, "Systematic design of a family of attack-resistant authentication protocols", *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 5, pp. 679–693, 1993.
3. U. Carlsen, "Cryptographic protocol flaws: know your enemy", In *Proc. 7th IEEE Computer Security Foundations Workshop*, pp. 192–200, 1994.
4. C.-C. Chang and T.-C. Wu, "Remote password authentication with smart cards", *IEE Proceedings E - Computers and Digital Techniques*, vol. 138, no. 3, pp. 165–168, 1991.
5. H.-Y. Chien, J.-K. Jan, and Y.-M. Tseng, "An efficient and practical solution to remote authentication: smart card", *Computers & Security*, vol. 21, no. 4, pp. 372–375, 2002.

6. W. Diffie, P. C. van Oorschot, and M. J. Wiener, "Authentication and authenticated key exchange", *Designs, Codes and Cryptography*, vol. 2, no. 2, pp. 107–125, 1992.
7. C.-L. Hsu, "Security of Chien et al.'s remote user authentication scheme using smart cards", *Computer Standards and Interfaces*, vol. 26, no. 3, pp. 167–169, 2004.
8. M.-S. Hwang and L.-H. Li, "A new remote user authentication scheme using smart cards", *IEEE Trans. on Consumer Electronics*, vol. 46, no. 1, pp. 28–30, 2000.
9. W.-C. Ku, S.-T. Chang, and M.-H. Chiang, "Weaknesses of a remote user authentication scheme using smart cards for multi-server architecture", *IEICE Trans. on Communications*, vol. E88-B, no. 8, pp. 3451–3454, 2005.
10. L. Lamport, "Password authentication with insecure communication", *Communications of the ACM*, vol. 24, no. 11, pp. 770–772, 1981.
11. S.-W. Lee, H.-S. Kim, and K.-Y. Yoo, "Improved efficient remote user authentication scheme using smart cards", *IEEE Trans. on Consumer Electronics*, vol. 50, no. 2, pp. 565–567, 2004.
12. G. Lowe, "An attack on the Needham-Schroeder public-key authentication protocol", *Information Processing Letters*, vol. 56, no. 3, pp. 131–133, 1995.
13. H.-M. Sun, "An efficient remote user authentication scheme using smart cards", *IEEE Trans. on Consumer Electronics*, vol. 46, no. 4, pp. 958–961, 2000.
14. W.-H. Yang and S.-P. Shieh, "Password authentication schemes with smart card", *Computers & Security*, vol. 18, no. 8, pp. 727–733, 1999.
15. E.-J. Yoon, W.-H. Kim, and K.-Y. Yoo, "Security enhancement for password authentication schemes with smart cards", In *Proc. 2nd International Conference on Trust, Privacy, and Security in Digital Business (TrustBus'05)*, LNCS 3592, pp. 90–99, 2005.
16. E.-J. Yoon, E.-K. Ryu, and K.-Y. Yoo, "An Improvement of Hwang-Lee-Tang's simple remote user authentication scheme", *Computers & Security*, vol. 24, no. 1, pp. 50–56, 2005.

Information Leakage and Capability Forgery in a Capability-Based Operating System Kernel

Dan Mossop and Ronald Pose

Faculty of Information Technology,
Monash University, Clayton,
Victoria, Australia 3800
{dgm, rdp}@csse.monash.edu.au

Abstract. The Password-Capability System has been designed as an operating system kernel suitable for general-purpose computing in a hostile environment. It has an access control mechanism based on password-capabilities, on top of which a confinement mechanism and a type management mechanism are layered. This paper studies the security of these mechanisms. We find that the mechanisms leak information which can be utilised by an attacker. Furthermore, we find that conditions placed on the generation of password-capabilities by the mechanisms enable the attacker to forge password-capabilities more efficiently than by exhaustive search. We show that all the discovered attacks can be prevented. This paves the way for the use of the mechanisms in a highly secure third-generation of the Password-Capability System.

Keywords: Password-capabilities, access control, security analysis.

1 Introduction

The Password-Capability System [1,2,3,4,5] operating system kernel was developed in the 1980s at Monash University, Australia. It introduced a new paradigm for implementing capability-based access control, the *password-capability*. Unlike tagged capabilities which require hardware support and segregated capabilities which must be kept separate from user data, password-capabilities can be both implemented on general-purpose hardware and freely mixed with other data. This can be done while retaining the naming, addressing and protection benefits associated with capabilities. This makes password-capabilities a promising paradigm for the development of secure operating systems. Password-capabilities were retained for access control in the Password-Capability System's successor, Walnut [2], developed in the 1990s. They have also formed the basis of access control in other systems including Mungi [6,7,8,9] and Opal [10].

The Password-Capability System and Walnut have shown that a number of powerful security features can be built on top of a password-capability-based access control mechanism, in a compact and efficient manner. Examples of such features are confinement and type management mechanisms. Work has begun to determine how this feature set can be improved and extended for a third-generation system. An earlier paper demonstrated that the Password-Capability

System can be extended to support arbitrary mandatory and discretionary security policies [11]. A covert channel analysis [12] has also been carried out on a formal model [1] of the system. Ongoing work suggests that the system can also support sophisticated controls on the distribution and revocation of capabilities, similar to those offered by (non-password) capability systems such as SPEEDOS [13,14] and Hydra [15].

In this paper, the security of the Password-Capability System's access control, type management and confinement mechanisms is analysed in detail.

We first look at the possibility of information leakage from the mechanisms. We find that by utilising combinations of the secrets on which the mechanisms are based, we can reconstruct useful information. We show that this information can be used to undermine the security of the system. We then present guidelines which, if adhered to, prevent information leakage and the attacks which exploit it.

We then consider attacks which attempt to discover capabilities by guess work. We find that regularities in the generation of the capability passwords enable their recovery faster than by exhaustive search. We show that the system can be modified in such a way that these guessing attacks are all rendered infeasible.

With the exception of a brute force attack on password capabilities, all attacks discussed in this paper apply only to the Password-Capability System and Walnut [2], and not to other password-capability based systems, such as Mungi [6,7,8,9] and Opal [10].

2 The Password-Capability System

In this section we introduce the Password-Capability System. We first give an overview of its main functionality. We then describe in detail two of its powerful security mechanisms: a type management mechanism and a confinement mechanism.

2.1 Overview

The Password-Capability System is an operating system with a global virtual address space, access to which is controlled by a mechanism utilizing password-capabilities. All entities such as data, files, processes, programs, in all such systems throughout the world, are considered to be objects. The virtual memory is divided into volumes, each of which is typically a storage device. Each object resides on a volume, and is uniquely identified in the system by an object name. The object name is comprised of a unique volume identifier and an object serial number unique on that volume.

Operations on objects are only permitted when a suitable password-capability is presented to the system. A capability comprises an object name and a randomly-chosen password. Each capability allows access to some aspects of an object. Aspects of objects that capabilities may give access to include the abilities to read or write a subset of the data in the object, the abilities to start,

stop, and send messages to objects that are processes, and the ability to derive new capabilities for the same object with a subset of the access rights.

Each object when created has a single master capability, from which derivative capabilities can be recursively created. The derivatives can never have more authority than their parents, and are destroyed if the parent is. Thus each object has associated with it, a singly-rooted tree of capabilities.

Capabilities are simply values; therefore the security of an object depends on the infeasibility of guessing any capability for the object.

Apart from being a store of data and potentially a process, all objects also act as stores of money. This enables an economic system to be created in which data and services can be traded, and resources managed in a familiar capitalist market model. Thus all objects are in effect bank accounts. Objects that are processes also contain some cash for spending on immediate needs such as CPU time, in the same way as people have ready cash for their immediate needs.

Processes have a *cashword* in which their cash is stored. They can send this money to other processes, or store it in other objects. The master capability indicates the object's total money in its *moneyword*. Derived capabilities' moneywords indicate withdrawal limits on the object's money. Thus the amount of money that can be withdrawn using a capability is the minimum of that of its moneyword, and that of all moneywords of its ancestors leading back to the master capability. The system periodically extracts an amount of money, proportional to the size of the object, from the moneyword of the object's master capability. This is a rental charge to pay for the system resources consumed. If rent can not be paid, the object will be deleted.

Processes can communicate with one another by passing short messages, or for larger messages, they can use an intermediate data object to which both have access. Message passing can be used to share capabilities to such an object.

2.2 Type Management and Confinement

The Password-Capability System supports a number of powerful security mechanisms. Two of these are its type management mechanism and its confinement mechanism.

Type Management. Sometimes we wish to be able to have some service process allow a client process to access to a data object, without the client being able to access the internal structure of the data object. An example of such a service is querying and updating a database. This can be achieved through the use of the Password-Capability System's type management mechanism.

For the purposes of type management, a capability's password is seen as comprising two parts, $p1$ and $p2$. When a type manager routine creates an object on behalf of a client process, the manager may *seal* the capability for the object by xoring the $p2$ field of the capability with a key k . The sealed capability is returned to the client, who must pass it back to the manager whenever requesting a further action on the instance. The manager can of course recover the original instance of the capability by xoring the $p2$ field of the sealed capability with k . The client, however, cannot access the instance.

Many systems can support type managers of this general sort. However, it is usually necessary for the client to call on the type manager for even the most routine management tasks. In our case it would seem that the client would have to call the type manager whenever the client wished to deposit funds in the object, or to create a derivative of the sealed capability. In fact, the Password-Capability System allows the client to use the sealed capability directly for these purposes.

If a client presents a sealed capability to the system, its original unsealed version can be found by the system by matching the object name and $p1$ fields alone. The system can then discover the seal key k as the xor of the true and sealed $p2$ fields. The system will accept the sealed capability as sufficient authority for any action within the rights of the true capability, except actions which could reveal or alter data in the data object being managed. If the requested action is that a derivative of the sealed capability is to be created, or for the sealed capability to be renamed, the system will seal the $p2$ field of the new capability password with k before returning it to the client. Thus a client may pay rent for or share use of the managed object without reference to the type manager, but any new capability created can be fully exercised only by the manager.

The first part of a capability's password, $p1$, is generated randomly (by the system) but with the condition that no other extant capability for the object has the same first password part. This is required by the type management mechanism, in order to identify capabilities as valid for certain operations based on the validity of their $p1$ value only. Note that while only some capabilities are ever validated in this way, all must abide by the rule of $p1$ uniqueness for the mechanism to work. The second part of the password is generated entirely at random. There is no requirement of uniqueness for this part. The original system specification recommended that the capability fields, v , s , $p1$ and $p2$ should all be 32 bits in length, giving a 128-bit capability.

As mentioned, the Password-Capability System will allow some operations to be carried out using capabilities for which only the first part of the password is correct. A simple attack against this scheme was described in [3], which involves guessing the first part of a password separately from the second part. To prevent this, the system was modified to return an apparently successful result on any operation where only a valid $p1$ is required. For instance, if a derivative is requested using a capability with an invalid $p1$, a 'fake' capability will be returned having a correct object name, but an entirely random (and, with high probability, invalid) password. This ensures that an attacker cannot determine the first part of a password separately from the second. This method of returning apparently successful results prevents that attack, but makes the first of our capability forgery attacks possible (Section 4.1).

Confinement. The Password-Capability System has a mechanism for addressing the confinement problem [16]. Each process in the Password-Capability System has an associated bit-string (equal in length to a capability password), called a *lockword*, which it cannot read. Any process having a suitable capability can modify the lockword of any other process by xor with an arbitrary value.

Whenever a process tries to use some capability which would enable it to communicate information (called an *alter* capability), the system first decrypts the capability's password by xor with the process's lockword before checking the capability's validity.

Initially a process has its lockword set to that of its creator. By default it is zero, in which case the xor has no effect and the process can use any capability it possesses. To confine a process X, some process Y can xor X's lockword to a chosen value, L. Now when X tries to use an alter capability, the capability password will be modified by the decryption process and the result will not be a valid capability, so the call will fail. X will still be able to use any *non-alter* capabilities it possesses.

Process X could try to guess its lockword. If it were able to do so then it could encrypt any alter capability in its possession using the guessed lockword, such that when it was presented to and decrypted by the system, the result would be the original capability and the call would succeed. This is prevented by having process Y set X's lockword to a value which is infeasible for X to guess. Lockwords are equal in length to capability passwords and these passwords are expected to be infeasible to guess. Hence guessing a randomly chosen lockword is also expected to be infeasible.

Sometimes it will be desirable to allow a confined process to use certain alter capabilities. Process Y can authorise alter capabilities for X's use by encrypting their passwords by xor with L (which it knows). These can then be passed to X and when used by X will be correctly decrypted by the system. Obviously process Y must not pass X an encrypted capability if X knows the unencrypted capability, since it will be able to derive its lockword from the pair.

We have described a single level of confinement. Because xor commutes, multiple levels of confinement can be implemented in a straight-forward manner.

3 Exploiting Leaked Information

Mechanisms, such as the ones described, must be protected against information leakage. We addressed the leakage of confidential information in our covert channel analysis [12]. Such information leakage has also been addressed in systems such as EROS [17]. Another form of information leakage is considered in this section: the leakage of access control information.

In actual use of the Password-Capability System, an attacker may legitimately come into possession of certain pieces of information, if they are authorised to do so. Examples are capability passwords, process lockwords and type management keys. The questions we wish to answer are: is it possible for the attacker to combine such secrets and recover information he has not been authorised to access? And if so, can we impose constraints on the behaviour of the mechanisms to prevent this?

It is easy to imagine certain situations in which an attacker could recover useful information. Suppose an attacker is given both an unauthorised version of an alter capability (i.e. an unmodified capability) and the authorised version

(i.e. the same capability, but with its password xored with some process's lockword). Then by xoring the password fields of these two capabilities together, the process's lockword will be recovered. This gives the attacker the ability to authorise other capabilities for the process and to break it out of its confinement, abilities it may not have been legitimately given. This situation was anticipated when the confinement mechanism was conceived. As a result it was recommended that only the authorised version, or the unauthorised version of a capability should be distributed, but not both. Should it be necessary to provide equivalent authority to the two versions, a newly derived capability equivalent to one of them should be used in its place.

3.1 Seal Transfer Attack

Is this condition sufficient to secure the mechanisms? Not necessarily. The following example describes a situation in which an attacker could gain access rights it should not have, by exploiting the type management mechanism. The situation is perhaps unlikely, but it plausible enough to cast doubt on the sufficiency of the above condition. Suppose we have a situation where we would like two separate type managers to control access to a single object. We may decide to create two capabilities giving access to two different parts of the object. If we want to give a process access to one part using the first type manager, we can seal the first capability's password (p_1) with the type management key (k_1) and pass it to the process. We may then decide that the process can use either type manager to access the second part of the object, so we could pass the process the two versions of the second capability (having password p_2), one sealed with the first type manager's key (k_1), and the other with the second type manager's key (k_2). Now that process possesses the values $p_1 \oplus k_1$, $p_2 \oplus k_1$ and $p_2 \oplus k_2$. It hasn't been given $p_1 \oplus k_2$, so it shouldn't be able to access the second area of the object through the second type manager. However, $(p_1 \oplus k_1) \oplus (p_2 \oplus k_1) \oplus (p_2 \oplus k_2) = p_1 \oplus k_2$. Hence the process can gain access that it should not be able to. Other variations of this attack may exist, so a suitable method of prevention should account for this. We give such a method next.

3.2 Preventing the Attack

Because both the type management mechanism and confinement mechanism use the xor operation, they can be used together, and can each consist of multiple layers. As a result, the encrypted and/or sealed capabilities which must be passed out to processes will have a password field of the following form:

$$p \oplus (k_1 \oplus k_2 \oplus \dots \oplus k_m) \oplus (l_1 \oplus l_2 \oplus \dots \oplus l_n),$$

with $m, n \geq 0$ and $m + n \geq 1$. Here k_i denotes the i th type manager key applied, and l_j the j th value applied to the process lockword.

In the two examples of possible information leakage given above, the password p was the same for more than one capability given out. This led to the possibility of the values being combined into a useful value, as $p \oplus p = 0$ (cancelling it out

and leaving only the key and lock parts) and $p \oplus p \oplus p = p$ (enabling the creation of a new capability with different keys and locks applied).

To prevent this we now propose that for each encrypted/sealed capability to be given out, the process giving it out first creates a new underlying password by deriving a new capability with equivalent access rights. This ensures that an attacker cannot get two encrypted/sealed capabilities with the same password (except with negligible probability). The passwords are chosen uniformly at random by the system and not disclosed to the attacker. Hence (the second parts of) the encrypted/sealed passwords made available to the attacker are in effect one-time pads. There is therefore no way for the attacker to gain any information about the keys and locks applied. Similarly, since the keys and locks are chosen uniformly at random there is no way for the attacker to gain any information about the passwords. The result is that the attacker cannot gain any information about the underlying secrets.

While the attacker cannot discover the secrets, it may be able to pull off a trick like that described earlier where the type management key was transferred from one capability to another. However, it is not possible for the attacker to extract the underlying password from any encrypted/sealed capability. If it tries to combine the passwords of multiple encrypted/sealed capabilities, the result will be some combination of each underlying password, along with the keys and locks. The combination of the underlying passwords will be the same as an actual password in the system with only negligible probability. Hence producing the combination is highly unlikely to give the attacker access rights beyond what he originally had.

As long as an equivalent capability is created each time one of the mechanisms is used no useful information will be leaked. Of course one should be wary of creating too many equivalent capabilities in that there is some overhead incurred. In such cases, it would be sensible to try to avoid the problem by using different data structuring.

4 Capability Forgery Attacks

In this section we present our capability forgery attacks against the system. A forgery attack is one which allows an attacker to determine a capability it has not been given by some other process. Typically the object name will be known and so only the password component need be determined for a successful forgery. Exhaustive search for a *specific* capability's password takes an average of $2^{L1+L2-1}$ guesses, with L1 and L2 denoting the lengths of the first and second parts of a capability password respectively. This is 2^{63} in the case of the original password length recommendations. The attacks we give here typically succeed with significantly fewer operations than this bound. We will also show that while these attacks demonstrate that the Password-Capability System (as originally designed) is insecure, the system can easily be modified to prevent them.

4.1 Forging Capabilities (I)

We now give the first of our forgery attacks. To test if some $p1$ value is a valid first part of a capability password we create $k(2^{L1/2})$ derivatives from a ‘fake’ capability containing it, where k is a small constant. If the value is correct then each of these derivatives will be valid capabilities and will therefore be mutually distinct in their $p1$ values. If our guessed value is incorrect then the $p1$ values returned will be random and by the birthday paradox we expect to find some repetition in these values, with high probability, assuming k is suitably chosen. The appearance of repetition is proof that our guess of $p1$ is incorrect and that we should try some other value. Suppose that there are 2^n extant capabilities for the object. An average of 2^{L1-n-1} guesses at $p1$ must be made before a valid $p1$ is located. Testing each $p1$ value takes approximately $k(2^{L1/2})$ guesses. Therefore, approximately $k(2^{(3L1/2)-n-1})$ operations must be carried out to identify the valid $p1$ value. Once a valid $p1$ value is found we can find the corresponding $p2$ with an average of 2^{L2-1} guesses. Hence the overall number of operations required by the attack is about $k(2^{(3L1/2)-n-1}) + 2^{L2-1}$. For the original recommendations of $L1 = L2 = 32$, this gives a total of number of operations of between about $k(2^{47})$ and $k(2^{33})$ (with normal situations approximating the higher of the two figures). This is a significant improvement over the previously best known means of forging capabilities, the brute force attack.

4.2 Forging Capabilities (II)

The second attack requires some capability to the object for which a forgery of some other capability is sought. We use the known capability to exhaustively create derivatives (of course, the capability must have the right to create derivatives). Each derivative uses up one of the finite number of unused $p1$ values. After 2^{L1} derivations we will have used up all possible $p1$ values. Those values for $p1$ which are not assigned to either our original capability or one of its derivatives must be assigned to some capability unknown to us. We can select one of these $p1$ values and then, with an average of 2^{L2-1} guesses, find the corresponding $p2$ value, giving us the full capability. This attack will typically produce a forgery with an average of $2^{L1} + 2^{L2-1}$ operations, or $3(2^{31})$ when using the original recommendations.

4.3 Preventing Forgery Attacks

We now consider how the system can be secured against the presented attacks. Both attacks require a capability’s $p2$ value (or equivalently, a type manager’s key) to be guessed. Since the $p2$ values are set entirely at random, there is no better way of discovering a $p2$ value than by exhaustive search of the possible values. This suggests that increasing $L2$ sufficiently will render all the attacks infeasible. Indeed increasing $L2$ will prevent any attack which involves finding a capability’s $p2$ by searching possible values.

The amount by which we should increase $L2$ depends on two factors: the number of operations the system can carry out in its lifetime and the level of confidence we require in the security of capabilities. At the time the system is implemented, an upper bound on the number of operations it will carry out in its lifetime can be estimated, based on the known processing power of the system (it is the system which is the bottleneck in the attacks), and its expected lifetime. Obviously upgrades should also be taken into consideration. There will likely be a very large margin of error, but setting the bound particularly high should accommodate this. This bound gives the maximum work factor an attacker can achieve since the attack operations must be carried out by the system on the attacker's behalf. We expect that a value of 2^{64} will be appropriate for most systems, unless their lifetime is expected to be particularly long, or they are (or may be upgraded to be) particularly powerful. Of course if an attacker can in fact get near to this upper bound, he may indeed succeed in attacks requiring that number of operations on average. It should be noted here that assuming an attacker can carry out at most 2^{64} operations implies that the brute force attack is feasible, since it requires at most 2^{64} operations. We are hence assuming a more powerful attacker than the original system did. If $p2$ is set to 64 bits in length we expect that an attacker may be able to succeed in the attack. By increasing this length by a single bit the attacker's chance of success will be at most $1/2$. Each extra bit halves his chance again. The number of bits we add in this way represents our confidence level. We suggest that a maximum probability of attack success of 2^{-32} should be adequate for most applications. This requires that we increase the length of $p2$ by 32 bits. This gives a total length for $p2$ of 96 bits.

We do not discuss it in detail here, but our analysis also uncovered a denial of service attack on the system. This attack requires that an attacker carry out 2^{L1} operations. To prevent it, we require the $p1$ field of capability passwords to be at least 64 bits in length.

In general, if a system will not be able to carry out more than 2^M operations and we want an attacker's maximum probability of success to be 2^{-C} then we should set $L1 = M$ bits and $L2 = M + C$ bits. We recommend that for the vast majority of systems, setting $L1 = 64$ bits and $L2 = 96$ bits should provide a high level of security for the foreseeable future. This gives us a total capability password length of 160 bits.

5 Conclusion

We are currently developing a highly secure third generation Password-Capability System. In preparation for the system, the paper studied the security of the current system's main security mechanisms. We found that the mechanisms leaked information which could be exploited by an attacker. We also found that an attacker could forge capabilities more efficiently than by exhaustive search. We showed that all the discovered attacks could be prevented.

References

1. Mossop, D., Pose, R.: Semantics of the Password-Capability System. In: Proceedings of the IADIS International Conference, Applied Computing 2005. Volume 1. (2005) 121–128
2. Castro, M.D.: The Walnut Kernel: A Password-Capability Based Operating System. PhD thesis, Monash University (1996)
3. Wallace, C.S., Pose, R.D.: Charging in a secure environment. Security and Persistence, Springer-Verlag (1990) 85–97
4. Anderson, M., Wallace, C.S.: Some comments on the implementation of capabilities. The Australian Computer Journal **20** (1988) 122–130
5. Anderson, M., Pose, R.D., Wallace, C.S.: A password-capability system. The Computer Journal **29** (1986) 1–8
6. Heiser, G., Elphinstone, K., Vochtelo, J., Russell, S., Liedtke, J.: The Mungi single-address-space operating system. Software Practice and Experience **28** (1998) 901–928
7. Vochtelo, J.: Design, Implementation and Performance of Protection in the Mungi Single-Address-Space Operating System. PhD thesis, University of NSW, Sydney 2052, Australia (1998)
8. Vochtelo, J., Elphinstone, K., Russell, S., Heiser, G.: Protection domain extensions in Mungi. In: Proceedings of the 5th IEEE International Workshop on Object Orientation in Operating Systems, Seattle, WA, USA (1996)
9. Vochtelo, J., Russell, S., Heiser, G.: Capability-based protection in the Mungi operating system. In: Proceedings of the 3rd IEEE International Workshop on Object Orientation in Operating Systems, Asheville, NC, USA (1993)
10. Chase, J.S., Baker-Harvey, M., Levy, H.M., Lazowska, E.D.: Opal: A single address space system for 64-bit architectures. In: Proceedings of the Third Workshop on Workstation Operating Systems, ACM Press, New York, NY, USA (1992) 80–85
11. Mossop, D., Pose, R.: Security models in the password-capability system. In: Proceedings of IEEE TenCon'05, Melbourne, Australia. (2005)
12. Mossop, D., Pose, R.: Covert channel analysis of the password-capability system. In: Proceedings of the Asia-Pacific Computer Architecture Conference (ACSAC 2005), Singapore. (2005)
13. Keedy, J.L., Espenlaub, K., Hellman, R., Pose, R.D.: SPEEDOS: How to achieve high security and understand it. In: Proceedings of CERT Conf. 2000, Omaha, Nebraska, USA (2000)
14. Espenlaub, K.: Design of the SPEEDOS Operating System Kernel. PhD thesis, The University of Ulm, Germany (2005)
15. Cohen, E., Jefferson, D.: Protection in the Hydra operating system. In: Proceedings of the Fifth ACM Symposium on Operating System Principles, ACM Press, New York, NY, USA (1975) 141–160
16. Lampson, B.W.: A note on the confinement problem. Communications of the ACM **16** (1973) 613–615
17. Shapiro, J.S., Hardy, N.: EROS: A principle-driven operating system from the ground up. IEEE Software, January/February (2002) 26–33

Reverse Engineering of Embedded Software Using Syntactic Pattern Recognition

Mike Fournigault¹, Pierre-Yvan Liardet², Yannick Teglia², Alain Trémeau³,
and Frédérique Robert-Inacio¹

¹ L2MP-ISEN, Place Georges Pompidou, F-83000 Toulon, France
mike.fournigault@l2mp.fr

<http://www.l2mp.fr/doct/fournigault.html>

² ST Microelectronics, 77 Avenue O. Perroy, F-13790 Rousset, France

³ LIGIV, 18 Rue du professeur Benoît Lauras F-42000 Saint-Etienne, France

Abstract. When a secure component executes sensitive operations, the information carried by the power consumption can be used to recover secret information. Many different techniques have been developed to recover this secret, but only few of them focus on the recovering of the executed code itself. Indeed, the code knowledge acquired through this step of Simple Power Analysis (SPA) can help to identify implementation weaknesses and to improve further kinds of attacks. In this paper we present a new approach improving the SPA based on a pattern recognition methodology, that can be used to automatically identify the processed instructions that leak through power consumption. We firstly process a geometrical classification with chosen instructions to enable the automatic identification of any sequence of instructions. Such an analysis is used to reverse general purpose code executions of a recent secure component.

Keywords: Power Analysis, Side Channel, Chip Instructions, Reverse Engineering, Pattern Recognition.

1 Introduction

The purpose of this paper is to study how pattern recognition techniques can be used to classify power signals representing secure component instructions. More precisely, we apply these techniques to a smart card. Kocher et. al. showed in [1] that power variations are correlated to both component instructions and manipulated data. In consequence, the global power consumption of a microprocessor leaks information about the operations it processes. Especially, when the component processes data encryption operations, this information can be used to recover secret information from the embedded cryptosystem [1,2,3,4,5].

Power analysis attacks generally work on power consumption traces and perform global statistical processing of those signals to discover secret leakage. For example, a Differential Power Analysis (DPA) tries to correlate via a selection function, hypothetic values of key bits on power signals. Differences between correlated signals and original signals create peaks when correct key bits are

guessed. This technique enables to recover the secret key, but the attacker needs to know the type of cryptographic algorithm, to formulate the selection function in DPA. As DPA, many attacks could be improved with further information about the processed code, in consequence code recognition is an added value to the attacker.

The Simple Power Analysis (SPA) introduced by Kocher et. al. in [1], shows that the sequence of instructions can be appreciated along the power traces.

To refine SPA, tokenizing power signals and identifying current signatures can become an interesting tool to perform power analysis attacks, as well as an interesting tool to provide the setup of other kinds of side channel attacks. It can help in processing signal synchronisation and in identifying specific macro code instructions in power signals. For example, automatic identification of the *S*-box execution part of a DES, enables to determine the corresponding time interval. It can be used to improve the efficiency of a DPA attack, by registering traces only during this interval. It can also help to detect the key points of an algorithm and then order a Differential Fault Analysis (DFA) attack on those key points. A first attempt in automatic code recognition was made on side channels in [6]. Other methods based on statistical approaches were also published that focus on dedicated unknown part, but with known structures [7]. The interest of pattern recognition methods was shown in [6]. We continue this direction, but with a method that is more able to counter-act some counter-measures and that is more able to deal with sequences of instruction signatures.

The aim of this paper is to outline how to classify code instructions extracted from power signals by using pattern recognition methods, and to show how such a classification paired with sequential pattern relations can be used to perform a SPA. The organisation of this paper is as follows. First, we describe the studied power signals. Then, the pattern recognition tool to compare elementary signatures is detailed. Next, we state how to construct, signature models, during a learning stage, to which input signatures are compared. Afterwards, we explain how pattern sequences and their grammatical analysis can help in macro instruction recognition. Finally we present a practical application.

2 Experiments and Power Signals Under Study

In [8], M.L. Akkar shows that the power consumption $P(I)$ of an instruction I can be separated from: the general power of the instruction, the power due to data in input and output of I , and the component due to previous instructions and manipulated data. Experimentally, we observed that previous instructions and data have no significant impact on the result. So, in our consumption model, we choose to neglect the consumption component due to the previous instruction. In our experiments, manipulated data are unknown. In addition, our method characterizes the consumption signatures by taking into account data influence. In consequence, we do not make any difference between data in input and data in output. So, we assume that $P(I)$ is defined by:

$$P(I) = P_{gen} \times \mathcal{N}_{gen} + P_{data} \times \mathcal{N}_{data} \quad (1)$$

where P_{gen} denotes the general power of the instruction and \mathcal{N}_{gen} refers to its variations, P_{data} refers to the power due to data and \mathcal{N}_{data} is the corresponding noise. With those considerations, it means that, we have to characterize the signature for various manipulated data to identify a component instruction from its power signature. So, we have to proceed to several executions of the same instruction with various random data when possible.

We consider for this study, a secured microcontroller for smart cards, more precisely a 8 bits CISC microcontroller with a Von Neumann architecture. The processor is run in stable clock mode with a frequency $f_c = 7$ MHz. The principal component of consumption remains synchronous with the clock signal. We observe that our method used to analyse power signals is self-adaptative to the clock frequency. In order to characterize the power consumption of each instruction according to the model of the equation 1, we programme a loop to execute N times the same instruction. Signals are recorded on the power supply contact with a current probe and an oscilloscope Tektronik T3032B, at the sampling frequency of 2GHz. Each instruction trace is composed of at least six hundred cycle samples, from which one hundred of cycle samples is representative of the instruction code.

In a first step, we need to define the set of instructions to be identified for real code. We consider general CPU statements like "load", "add", "and", "or" and "multiply". We refer to the set of instructions as $\mathcal{I} = \{I_i, i \in [1, n]\}$ where I_i denotes a precise instruction. By precise we mean an instruction with an addressing mode. The data set chosen for the k^{th} instruction execution is noted $D_k, k \in [1, N]$, and $I_i^k = I_i(D_k)$ refers to the k^{th} execution of the instruction I_i with data D_k . $P_i^k, k \in [1, N]$, denotes the current signature of I_i^k .

In order to identify power traces according to the consumption model of the equation 1, we proceed in two stages. The first stage consists in learning characteristics of instruction signatures in order to get some specific knowledges. We will refer to this step as the learning stage. According to equation 1, power signatures of an instruction was characterized without separating the contribution of the instruction and the contribution of the data. It finally leads, for each instruction I_i , to the choice of a set of signature prototypes that constitutes a reference database of signatures statistically representatives of I_i . The second stage consists in identifying a power signature P_x given in input, by searching for a signature prototype that is similar to P_x . Each signature database of each instruction characterized, is scanned in order to find the signature prototype of I_i , that is the most similar to P_x . If the similarity degree between both is higher than a threshold value, P_x is said to be characteristic of the instruction I_i . Both learning stage and identification stage use a measure of similarity between signatures. We describe hereafter how such a similarity is measured.

3 Elementary Pattern Recognition Scheme

Both learning stage and identification stage need a tool to compare signatures.

As an instruction I_i takes a constant number (noted R_i) of cycles to execute, a signature P_i^k is decomposed, using a gaussian derivative wavelet transform [9],

in subparts corresponding to cycle signatures, where R_i are examined (see fig. 1). In our scheme, each cycle signature is disassembled in significant peaks. Each peak is characterized individually, it allows to take into account data influence on the signature. From each significant peak of the cycle signature $P_{i,c}^k$, a set of shapes $S_{c,j}$ is constructed by considering the subgraph (see fig. 1). According to this decomposition, matching two cycle signatures $P_{i,c}$ and $P_{i,d}$ consists in comparing every shape of $P_{i,c}$ to every shape of $P_{i,d}$.

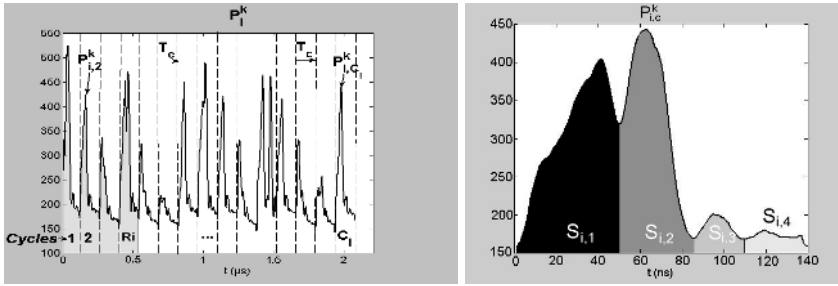


Fig. 1. Left: decomposition of P_i^k in cycle signatures. Right: decomposition of a cycle signature in elementary shapes.

In the learning stage, shapes are used to build shape prototypes. In the identification stage, input shapes are compared to learned prototype shapes.

These two stages are based on a pattern recognition tool defined by F. Robert in [10]. This tool is a shape parameter, that defines a similarity degree between two shapes, the shape under study and a reference shape. The Robert’s parameter is a bounded measure that enables to compare shapes according to some geometrical features. An important fact for our application is that this parameter is invariant under translation and scaling. This allows to take into account some counter-measures running on the studied component that modifies the magnitude of consumption peaks.

This parameter considers two convex shapes, on the one hand X , is the shape under study and, on the other hand A is the reference shape. In order to compute the Robert’s parameter, we search for the smallest homothetic set of A circumscribed to X , $A_X = \lambda_A(X).A$, where $\lambda_A(X)$ denotes the scale ratio to apply to A . Then, the smallest homothetic set of X circumscribed to A_X , $X_{A_X} = \lambda_X(A_X).X$, is computed. In this way, the two shapes A and X are compared with the Robert’s shape parameter, as follows:

$$p_{A_X}(X) = \frac{1}{\lambda_X(A_X)} \frac{\mu(X)}{\mu(A_X)} \tag{2}$$

where $\mu(X)$ and $\mu(A_X)$ refer respectively to the areas of X and A_X .

The Robert’s parameter is presented in [10] for convex shapes. Since in our application, the elementary signatures are often quasi convex and also the corresponding patterns S_j , it would be quite restrictive to deal only with convex

shapes. A way to compute this parameter for non convex shapes, is to proceed as explained in [11].

We note $p(S_{c,j}, S_{d,l})$ the similarity measure between the shape $S_{c,j}$ extracted from $P_{i,c}$ and the shape $S_{d,l}$ extracted from $P_{i,d}$. The similarity measure between two cycle signatures $P_{i,c}$ and $P_{i,d}$, $M(P_{i,c}, P_{i,d})$ is:

$$M(P_{i,c}, P_{i,d}) = \sum_j \max_l (p(S_{c,j}, S_{d,l})) \quad (3)$$

4 Learning Stage and Current Signature Sequences in Grammatical Formulation

The learning stage is the operation of determining for each instruction I_i , the set of signature prototypes that are characteristic of its power consumption, regardless of the manipulated data. Each instruction is assigned to several signature prototypes $\overline{P}_i^1, \dots, \overline{P}_i^q$. We note $\overline{\mathcal{P}}_i$, the set of all signature prototypes of I_i .

As mentioned in section 3, the number of cycles to execute I_i , R_i , is known during the learning stage. So, each prototype of $\overline{\mathcal{P}}_i$ is constructed from R_i prototypes $\overline{P}_{i,c}$ of cycle signatures. The learning stage begins with the choice of cycle signature prototypes $\overline{P}_{i,c}$ that are representatives of those possibles for I_i . It continues with the construction of each prototype of instruction signature \overline{P}_i^q from cycle prototypes.

A cycle prototype $\overline{P}_{i,c}$ is said to be characteristic of n samples of cycle signature (regardless to the data influence), if it is similar, according to a threshold value T , to n samples of cycle signature, where n is as large as possible with respect to the threshold T . The prototype $\overline{P}_{i,c}$ is chosen to be

$$\overline{P}_{i,c} = P_{i,c}, c \in [1, n], \forall d \in [1, n], M(P_{i,c}, P_{i,d}) \geq T$$

Maximising n for each prototype results in minimising the number of prototypes $\overline{P}_{i,c}$ required to characterize an instruction. In consequence, it results in minimising the matching complexity during the identification stage. Finally, the set of prototypes of instruction signatures are learned to be all possible sequences of cycle prototypes encountered.

We consider the problem of identifying an instruction signature in a power trace, as matching a string with another string of a reference language.

During the learning stage, the language $L(G_i)$ associated to each instruction I_i is the set $\overline{\mathcal{P}}_i$ of instruction signature prototypes. The grammar G_i is the set of rules to derive each \overline{P}_i^q in R_i cycles prototypes $\overline{P}_{i,1}^q \wedge \dots \wedge \overline{P}_{i,R_i}^q$, where \wedge is the concatenation operator, and finally, each cycle prototype $\overline{P}_{i,c}^q$ in elementary shapes $\overline{S}_{c,j}^q$. The registration of the set $\overline{\mathcal{P}}_i$ of prototypes enables string recognition as direct string matching, without any parser tool.

In this way, the identification of an instruction through its power signal is equivalent to the simple matching of its string representation to some reference strings of a database constructed during the learning stage. An input signature

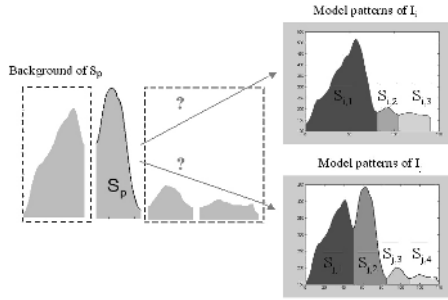


Fig. 2. Recognition of a cycle signature according to syntactic methods. The elementary pattern S_p is discriminant in the identification of the input cycle signature.

P_x is similar to a signature prototype $\overline{P_i^q}$, if for each cycle signature $P_{x,c}$ of P_x and each cycle signature $\overline{P_{i,c}^q}$ of $\overline{P_i^q}$, $M(P_{x,c}, \overline{P_{i,c}^q}) \geq T$. Each signature database of each instruction characterized, is scanned in order to find the signature prototype of I_i , that is the most similar to P_x . This syntactic analysis is generalized to instruction sequences, and allows us to recognize macro code instructions and their sequential execution. When combined with previous pattern matching, this syntactic analysis provides a tool to reverse code through power analysis.

In figure 2, we illustrate the syntactic recognition process with the example of the identification of a cycle signature.

5 Experiment Results on a Recent Secure Component

In this section, we present some application results obtained for instruction identification of a recent secure component. This component runs some counter-measures and especially a magnitude counter-measure that occurs randomly on consumption peaks during the instruction execution. This component also embeds a phase jitter counter-measure that had been stopped for our experiments.

In order to give these application results, all instructions that we want to identify, have been previously characterized during a learning stage, and so, that all signature prototypes are availables in the registered databases.

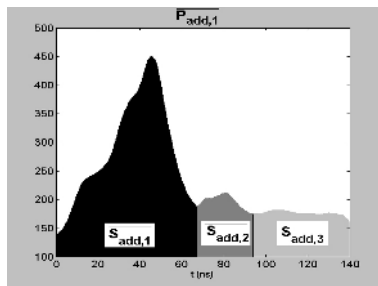


Fig. 3. Pattern models extracted from $\overline{P_{add,1}^1}$

We first illustrate the identification of an instruction "add". We begin to illustrate this example with the matching of a cycle signature, and then the complete instruction signature is presented. The signature model of the first instruction cycle is extracted from the prototype $\overline{P_{add}^1}$ and is noted $\overline{P_{add,1}^1}$. Three pattern models are separated: $\overline{S_{add,1}^1}$, $\overline{S_{add,2}^1}$, $\overline{S_{add,3}^1}$.

We describe the result of matching two signature samples of I_{add} , $\overline{P_{add}^1}$ and $\overline{P_{add}^2}$ with $\overline{P_{add}^1}$. The first cycle signatures of those two samples are noted respectively $\overline{P_{add,1}^1}$ and $\overline{P_{add,1}^2}$. We give on fig. 3 the three pattern models computed from $\overline{P_{add,1}^1}$, and those of $\overline{P_{add,1}^1}$ and $\overline{P_{add,1}^2}$ are given on fig. 4.

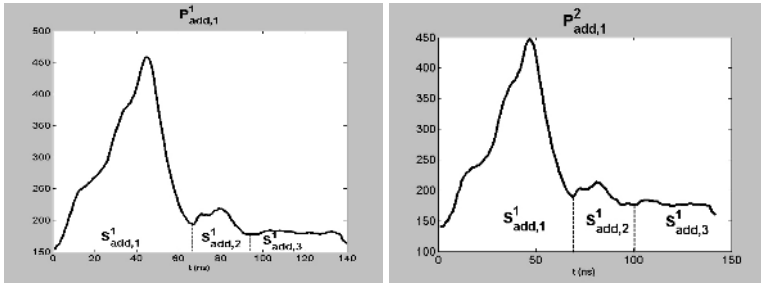


Fig. 4. Left: $\overline{P_{add,1}^1}$ and its patterns. Right: $\overline{P_{add,1}^2}$ and its patterns.

In order to proceed to the identification of power signals $\overline{P_{add,1}^1}$ and $\overline{P_{add,2}^2}$, all input patterns are compared to all patterns of $\overline{P_{add,1}^1}$. In tab. 1, we have reported the shape parameter value between $\overline{S_{add,j}}$ and $\overline{S_{add,j}^k}$, $j = \{1, \dots, 3\}$ and $k = \{1, 2\}$, for the shapes of fig. 4 according to model patterns of fig. 3.

Table 1. Best scores for the comparison of $\overline{S_{add,j}^k}$ and $\overline{S_{add,j}}$

	$\overline{S_{add,1}}/\overline{S_{add,1}^k}$	$\overline{S_{add,2}}/\overline{S_{add,2}^k}$	$\overline{S_{add,3}}/\overline{S_{add,3}^k}$
$\overline{P_{add,1}^1}$	0.48	0.35	0.37
$\overline{P_{add,1}^2}$	0.46	0.71	0.54

Although values of tab. 2 seem to be quite low, they are ten times higher than values obtained when comparing $\overline{S_{add,j}^k} / \overline{S_{add,o}}$ with $o \in [1, 3], j \in [1, 3], o \neq j$. It leads in similarity measures $M(\overline{P_{add,1}^1}, \overline{P_{add,1}^1}) = 1.2$ and $M(\overline{P_{add,1}^2}, \overline{P_{add,1}^1}) = 1.71$. The threshold value used is $T = 1$, it enables to say that the cycle signatures $\overline{P_{add,1}^1}$ and $\overline{P_{add,1}^2}$ are similar to the cycle prototype $\overline{P_{add,1}^1}$.

We now consider the entire signature $\overline{P_{add}^1}$ and $\overline{P_{add}^2}$. In this example, $\overline{P_{add}^1}$ matches the prototype signature $\overline{P_{add}^1}$, with respect to the minimum similarity value T and where no other signature prototype of any $I_i, i \neq add$ gives better results. Because of its second cycle signature, $\overline{P_{add}^2}$ does not match $\overline{P_{add}^1}$, but $\overline{P_{add}^2}$

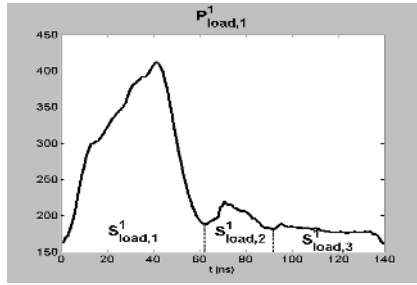


Fig. 5. Example of a cycle signature of a load instruction

matches another prototype of "add" signature, $\overline{P^4_{add}}$ giving the highest similarity measures with P^2_{add} than any other prototype of any instruction. This example shows that our method enables to identify the instruction "add" through its power signatures. In our experiments, we identify the instruction "add" with at least 75% of success on the secure component tested.

We now illustrate this recognition scheme with an input signature corresponding to the instruction "load" executed with random data. This sample is noted P^1_{load} . We begin to describe this example with the matching of one cycle signature, noted $P^1_{load,c}$ (see figure 5). We present the comparison of $P^1_{load,c}$ to the cycle prototype of the instruction "add" that gives the best scores, $\overline{P_{add,1}}$.

In tab. 2, we have reported the shape parameter values between $S_{load,j}$ and $\overline{S_{add,j}}$, $j = \{1, \dots, 3\}$, that correspond to best scores of comparisons $S_{load,j}/\overline{S_{add,o}}$, $j \in [1, 3]$, $o \in [1, 3]$. It leads in the similarity measure $M(P^1_{load,1}, \overline{P^1_{add,1}}) = 0.85 < T$. The best scores for $S_{load,2}$ and $S_{load,3}$ are equivalent to those of $S^k_{add,2}$ and $S^k_{add,3}$. But the score for $S^k_{add,1}$ is more than 2 times better than the score for $S_{load,1}$. From this example we verify that, according to the threshold value $T = 1$, shape parameter values are discriminant enough to conclude that $P^1_{load,1}$ is not similar to $\overline{P^1_{add,1}}$.

We now consider the entire signature P^1_{load} to signature prototypes of the instruction "add". The tested instruction "load" takes 3 cycles to execute, and so, the first three cycle signatures of P^1_{load} are tested. The string of P^1_{load} cannot be matched to any string prototype of the instruction "add", and we verify for this example that the signature P^1_{load} is not characteristic of the instruction "add".

Finally, we give on figure 6, the application of identifying subparts of an input signature. It was successfully identified as the sequence of two different "load" instruction signatures, followed by an "XOR" instruction signature.

Table 2. Best scores for the comparison of $S^1_{load,j}$ to $\overline{S_{add,o}}$

	$S_{add,1}/S^1_{load,1}$	$S_{add,2}/S^1_{load,2}$	$S_{add,3}/S^1_{load,3}$
$P^1_{load,1}$	0.19	0.36	0.30

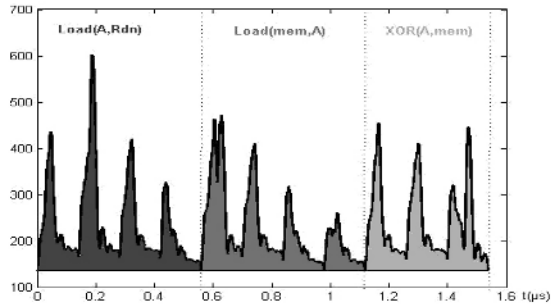


Fig. 6. Identification of subparts of an input signature

6 Conclusion

In this paper, we have shown that pattern recognition methods could automatically identify instructions through power signals of a recent secured smart card component. This process of instruction identification needs two steps: a characterization step to produce signature models and a recognition step to compare input signatures to signature models. Tokenizing instruction signatures in cycle signatures, and then, in elementary patterns, allows to perform a local analysis. This local analysis powered by a shape parameter and a syntactic analysis enables to automatically identify precise subparts of an instruction signature, such that a cycle signature and then the complete instruction signature. According to the results of our experiments, our pattern recognition scheme enables to recognize 75% in the worst case, and 81% in the average case, of tested instruction signatures, showing that this is an interesting tool to reverse code instruction from power signals. Most of the identification failures encountered are due to counter-measures that were running. Although our method is really efficient for some counter-measures like amplitude counter-measure, it does not work when the jitter phase counter-measure is activated. Finally, let us outline that our method needs to be applied on an opened component on which it is possible to execute specific instructions, in order to learn signature prototypes.

In future work, it can become interesting to test a different syntactic analysis scheme like regular grammar analysis, in order to analyse how much the previous instruction and elementary statement can influence the power consumption.

References

1. Kocher, P., Jaffe, J., Jun, B.: Differential Power Analysis. *Lecture Notes in Computer Science* **1666** (1999) 388–297
2. Fahn, P., Pearson, P.: IPA: A New Class of Power Attacks. *Lecture Notes in Computer Science* **1717** (1999) 173–186
3. Clavier, C., Coron, J.S., Dabbous, N.: Differential Power Analysis in the Presence of Hardware Countermeasures. *Lecture Notes in Computer Science* **1965** (2000) 252–263

4. Berna Ors, S., Gurkaynak, F., Oswald, E., Preneel, B.: Power-Analysis Attack on an ASCIC AES implementation. IEEE ITCC 04 proceedings **2** (2004) 546
5. Mangard, S.: A Simple Power-Analysis (SPA) attack on implementations of the AES key expansion. ICISC 02 Proceedings, Lecture Notes in Computer Science **2587** (2002)
6. Quisquater, J.J., Samyde, D.: Automatic Code Recognition for Smartcards Using a Kohonen Neural Network. Proceedings of the Fifth Smart Card Research and Advanced Application Conference (CARDIS '02). San Jose, USA, november (2002)
7. Clavier, C.: Side Channel Analysis for Reverse Engineering (SCARE) - An Improved Attack Against a Secret A3/A8 GSM Algorithm. Cryptology ePrint Archive, <http://eprint.iacr.org/>, Report 2004/049, (2004)
8. Akkar, M.L.: Attaques et méthodes de protections de systèmes cryptographiques embarqués. Doctor Thesis, Versailles University (2004)
9. Bigot, J.: A scale-space approach to landmark detection. Technical Report, TR2046, PAI (Interuniversity Attraction Pole network). (2002)
10. Robert, F.: Shape studies based on the circumscribed disk algorithm. IEEE CESA 98 proceedings, IEEE-IMACS, Hammamet, Tunisia, 1-4 april. (1998)
11. Fournigault, M., Trémeau, A., Robert-Inacio, F.: Characteristic centre point for quasi-convex shapes. 9th European Congress on Stereology and Image Analysis proceedings (ECSIA), Zakopane, Poland, 10-13 May. **2** (2005) 299–304

Disaster Coverable PKI Model Utilizing the Existing PKI Structure^{*}

Bo Man Kim¹, Kyu Young Choi¹, and Dong Hoon Lee²

Center for Information Security Technologies(CIST),
Korea University, 1, 5-Ka, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
{bmkim, young}@cist.korea.ac.kr,
donghlee@korea.ac.kr

Abstract. A Public Key Infrastructure (PKI) using a certificate has already been widely used in a variety of fields for the provision of security service. Accordingly a Certification Authority (CA) that issues a certificate must securely construct a Certification Authority System and manage it. It is significantly important for a CA to ensure its service to continue to operate properly by preparing for any disaster caused by a CA's private key compromise no matter what the cause is.

In this paper, we provide the definitions of PKI disaster recovery and PKI business continuity, which are more clear and specific than ever before. We also present three requirements for a PKI model preparing for a disaster. Then we propose a PKI model that ensures business continuity in the event of a disaster in which a CA key is exposed. It is easily applied to the existing PKI structure. We stress that the proposed PKI model in this paper is the first to ensure both applicability to the existing models and business continuity in the event of a disaster.

Keywords: Public key infrastructure, PKI model, business continuity, forward secure signature scheme.

1 Introduction

BACKGROUND. Over the past several years, e-commerce using the internet has been rapidly growing in various ways such as internet banking, cyber stock trading, electronic payments, and other web services. However, there are a wide variety of inherent risks in e-commerce and we are inevitably exposed to them. With the expansion of e-commerce, the role of a PKI, namely, to ensure security and reliability of e-commerce is growing in importance.

The PKI is one of the most critical techniques to support e-commerce by ensuring authentication, integrity, non-repudiation, confidentiality, and access control. The distinguishing feature of a PKI is the use of a user's digital certificate issued by a CA. A CA in PKI issues a certificate which is digitally signed using a CA's secret key to messages specifying a user and the corresponding

^{*} This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

public key. Thus, a CA's secret key compromise is directly connected to a PKI disaster. Once the certification service is discontinued by a CA's key compromise, it will result in considerable degradation of the overall level of reliability and finally the fatal impact to the overall PKI.

A large amount of corporate attention has been rightly paid to Business Continuity Planning (BCP) as an important means of a disaster preparation since the September 11th attack in the United States. The purpose of BCP is to provide for the continuation of critical business functions and to minimize the effect of disruptions in the event of an unanticipated disaster. In spite of the importance of BCP in PKI for the stable PKI operation in the event of a disaster, no previous work has fully addressed this issue.

RELATED WORK. As mentioned above, there are few works closely related to our study. We will however summarize some of works that have inspired us. In 2003, Koga et al. presented the distributed trust CA model [4] in which the length of certification path is shortened by using Forward Secure Signature (FSS) and Key-Insulated Signature (KIS). This distributed CA trust model is not a primary solution to ensure business continuity within PKI when a disaster happens, but it can mitigate the damage caused by a CA's secret key exposure. Le et al. suggested the hierarchical CA trust model [6] which converts the existing key-insulated methods to a hierarchical scheme. This model shortens the verification path using KIS that many secret keys share one fixed public key and minimize the damage by a CA's secret key exposure. However, the model in [4,6] is not designed primarily for business continuity within PKI. Tzvetkav proposed the disaster coverable PKI model [8] which focused on PKI business continuity based on the Majority Trust principle. This model extends the PKI with a reliable and resistant mechanism against a CA's key compromise because the trust responsibility is distributed to multiple trusted CAs. The existing certificates can be continually used after a CA's secret key exposure. The disadvantage of this model, however, is the inefficiency occurring with the existence of multiple CAs.

CONTRIBUTION. In this article, we first define the terms of PKI disaster recovery and PKI business continuity that we will use in our proposed model. To the best of our knowledge, no other previous work has provided these definitions. We set three requirements for a desirable PKI model that ensures PKI business continuity. Finally, we propose a disaster coverable PKI model to mitigate the damage and to ensure business continuity in the event of a PKI disaster, namely, a CA's secret key compromise. It is called *disaster coverable PKI model* in this paper. The major advantage of our proposed model is that it can be easily applied to the existing PKI structures. We concentrate on the fact the our proposed model in this paper is the first to provide both applicability to the existing models and business continuity in the face of a disaster.

ORGANIZATION. In section 2, we provide the definitions of PKI disaster recovery and PKI business continuity. In section 3, we propose the PKI model that ensures business continuity in the event of a CA's secret key exposure. In section 4, we analyze our proposed model in detail, to gain deeper understanding of our study. Summary and concluding remarks are made in section 5.

2 Disaster Recovery and Business Continuity in PKI

2.1 Disaster Recovery Planning and Business Continuity Planning

Both Disaster Recovery Planning (DRP) and BCP are a series of procedures that respond efficiently and timely to disaster contingencies that disrupt the normal operations and have fatal impacts to the parties involved. In brief, they describe how an organization will deal with potential disasters. The concepts of BCP and DRP are often used confusedly in many cases. In this paper, we regard DRP as a subset of the broader concept of BCP. While BCP is concerned primarily with the continuation of business when a disaster happens, DRP is about getting back to normal after a disaster. We introduce the concept of PKI disaster recovery and PKI business continuity as follows:

2.2 PKI Disaster Recovery and PKI Business Continuity

The concepts of DRP and BCP in PKI have to be understood in light of a CA that provides certification services and a user that receives that services. A CA in which a disaster occurs is the place where DRP within PKI is required in the face of a disaster. A user that uses certification services becomes a beneficiary of BCP when a disaster happens. In the case that a CA is damaged by any disaster, the DRP and BCP within PKI are defined as following;

PKI disaster recovery planning: It enables a CA to perform the same functions as prior to a disaster by restoration of all systems and resources to full, normal operational status.

PKI business continuity planning: It enables a user to continue to use the same certification services as prior to a disaster using the existing certificate without reissuing a new certificate.

2.3 Requirements for a Disaster Coverable PKI Model

To study a desirable PKI model that ensures business continuity, both security and efficiency should be significantly considered. We present the three requirements that will affect the design of our model.

Independence of CA's key generation: In a hierarchical structure, the root CA issues certificates to their subordinates and certifies their subordinate CA's public keys. The root CA also has a responsibility for supervising and monitoring the subordinate CAs. However, the root CA should not be allowed to engage in CA's key generating. A private key and a public key of a CA should be separately issued and managed independently by each CA itself. Also, the information of a CA's secret key should not be shared with other CAs.

Efficiency of CA's key management: If a CA has multiple private and public keys, it tends to incur huge management expenses and complexity. Thus the CA should limited to as few keys as possible.

Applicability to the existing PKI structure: It is very inefficient to set up a new PKI model because plenty of time and expense are required. The existing PKI structure should be maintained.

3 Our Proposed Disaster Coverable PKI Model

In this section, we discuss a forward-secure signature scheme (FSS) and then propose our PKI model that ensures business continuity using FSS. Prior to it, we provide an overview of FSS adopted in our model.

3.1 An Overview of FSS

A lot of the digital signature schemes have been proposed, but they have not provided any security guarantees in the case of a secret key being exposed. In practice, the greatest threat against the use a digital signature is a secret key exposure. Once a secret key is compromised, any message can be forged. In order to mitigate the damage caused by a secret key compromise, the concept of FSS was initially proposed by Anderson [2] and formalized by Bellare and Miner [3]. Since then, more works on FSS have been proposed [1,5,7]. The basic idea of FSS is to extend a conventional digital signature algorithm with a key update algorithm. While the public key stays fixed, the secret key can be changed at regular intervals so as to provide a forward security property. Compromise of the current secret key does not enable an adversary to forge signatures pertaining to the past. This FSS can be very useful to mitigate the damage caused by a key exposure without distribution of keys. We propose the PKI model that ensures business continuity using FSS.

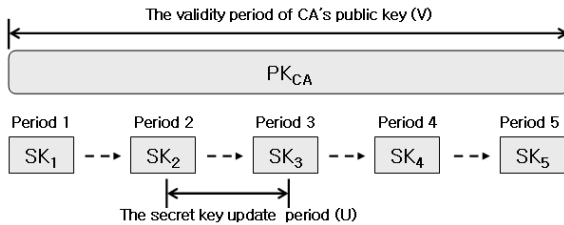


Fig. 1. The secret key update process in FSS

3.2 Introduction of Our Proposal PKI Model

Our proposed model is based on the PKI hierarchial structure and CAs take FSS as a signature algorithm when issuing a certificate. The proposed PKI model is as follows;

In practice, many of the existing PKI models have a hierarchical structure. In this paper, we propose a disaster coverable PKI model which is based on a

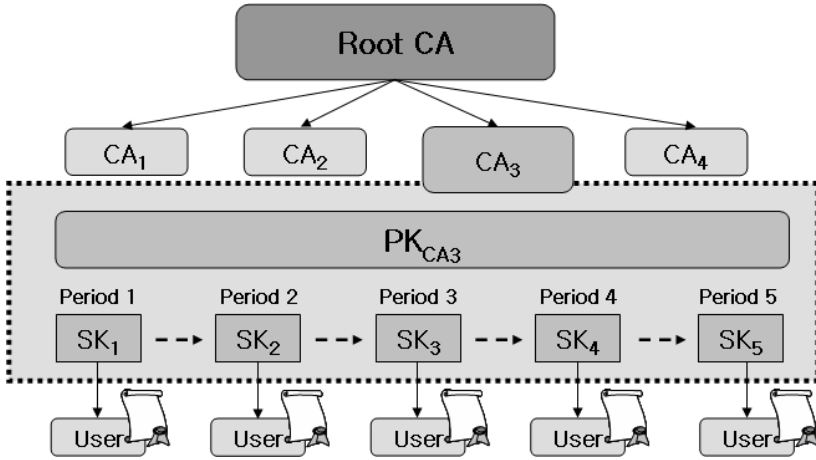


Fig. 2. Our proposed PKI model that provides business continuity

hierarchical organization, however, the CA mechanism using FSS produced in our model can be applied to different kinds of PKI structures.

Initialize and Generate the key pair: A CA performs the initialization step of the FSS scheme. Then, the CA determines a secret key’s update period (U). The total number of time periods (T) is produced using the validity period (V) of a CA’s public key certificate as follows;

$$T = \frac{V}{U}$$

For example, suppose that the validity period (V) of a user’s public key certificate is a period of 5 years and a CA secret key update period (U) is one month. The total number of the update of time periods (T) is 60. T is closely related to PKI business continuity in the event of a disaster. The range that guarantees business continuity gets larger as the total update number of a CA’s secret key (T) gets bigger.

After performing the initialization step, a CA generates a public key (PK) and an initial secret key (SK_0) according to the key generating algorithm of FSS.

Update the secret key: A CA evolves its secret key every period. Thus in each period, a CA produces signatures using a different secret key. The secret key update is performed at the end of each time period, at the same time of key update, the CA also deletes the previous secret key.

Issue the certificate: A certificate is issued (i.e. digitally signed) by a CA’s secret key. When a certificate is issued, FSS is used as a signature algorithm and the secret key that is used in signing only belongs to the current time period.

Verify the certificate: While a secret key evolves over time, the CA’s public key stays unchanged. Thus the certificate verification path is the same as the

current PKI model. Notice that our proposed model has the same efficiency as the existing model has.

4 Analysis of the Proposed Disaster Coverable PKI Model

4.1 Ensuring Business Continuity

The most important aspect of PKI business continuity is that a user can continue to use the same certificate issued prior to a disaster after a CA secret key exposure. In our model, business continuity is guaranteed because all the certificates that were issued prior to a disaster can be used continually during a disaster recovery period or even after it. Suppose that a secret key (SK_j) is exposed in period j , any information from the previous keys (SK_1, \dots, SK_{j-1}) will not be leaked since the secret key evolves under FSS. The user can continue to use the certificate that was issued prior to period j .

4.2 Comparison of Our Proposed Model and the Existing PKI Models

The PKI model [8] by Tzvetkov was proposed primarily for ensuring business continuity in the event of a CA's disaster. In the following table, we compare our proposed model with the existing PKI models [8] by using the requirements of the desirable PKI model preparing for a disaster discussed in section 2.

Table 1. Comparison of the proposed model and the existing PKI model

	Independency of CA's key generation	Efficiency of CA key management	Applicability to the existing PKI	Other advantages
Tzvetkov's model [8]	satisfactory	unsatisfactory	unsatisfactory	ensuring business continuity
Our proposal model	satisfactory	satisfactory	satisfactory	ensuring business continuity

Table 1 shows the competitive advantages of our model. As seen in Table 1, our proposed model is an efficient PKI model that ensures business continuity and satisfies the all three requirements that we presented earlier.

4.3 Relationship Between a CA Secret Key Update Period and the Range of PKI Business Continuity in the Event of a Disaster

In order to demonstrate the relationship between a CA secret key update period and the range of PKI business continuity in the event of a disaster in our proposed model, we define the denotation as follows:

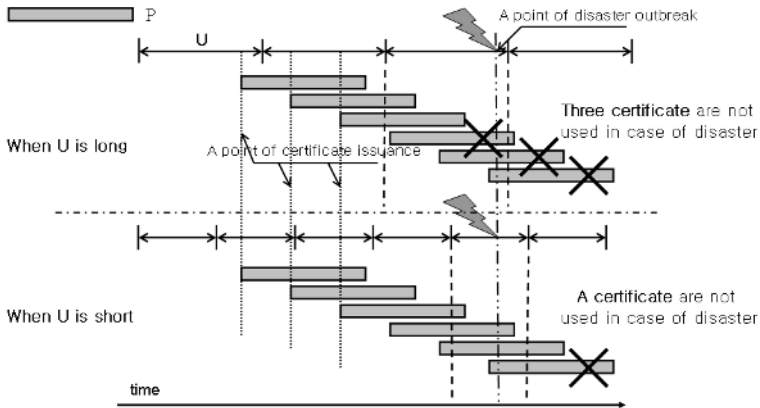


Fig. 3. Relationship between the CA secret key update period and the range of PKI business continuity

- *P*: the validity period of a user’s public key certificate issued by CA.
- *V*: the validity period of a CA public key.
- *U*: CA secret key update period ($U \ll P$).
- *T*: the total number of CA secret key updates.
- *BC*: the guarantee range of business continuity in the event of a disaster.
- *TotalCert*: the total number of the certificates issued by CA during *V*.
- *PeriodCert*: the average number of certificates issued in each time period.
- *RevCert*: the average number of certificates revoked in the event of a disaster.

As indicated above, the CA secret key update period (*U*) must be smaller than the validity period of a user’s public key certificate (*P*). In Figure 3, as the difference between *U* and *P* gets larger, the average number of certificates revoked in the event of a disaster (*RevCert*) gets smaller. Based on this, *T* and the average number of certificates issued in each time period (*PeriodCert*) can be produced as follows;

$$T = \frac{V}{U}$$

$$PeriodCert = TotalCert \times \frac{1}{T} = TotalCert \times \frac{U}{V}$$

Suppose that a CA perceives the disaster right after it occurs and the CA then discontinues all certification services that use a certificate. The average number of certificates revoked in the event of a disaster (*RevCert*) does not exceed the average number of certificates issued in each time period. This fact is illustrated in a simple function below.

$$RevCert \leq PeriodCert$$

If the guarantee range of business continuity (BC) in the event of a disaster defines the number of the valid certificates in the case of a CA disaster, the value of BC is yielded as follows:

$$BC = TotalCert - RevCert \geq TotalCert - PeriodCert$$

As seen above, as the value of $RevCert$ (or $PeriodCert$) gets smaller, the value of BC gets larger. To make the value of $RevCert$ small, the value of T must get larger. If CA has a shorter period of the secret key update (U), the total number of CA secret key updates (T) gets bigger. As a result of it, the guarantee range of business continuity (BC) gets larger. Thus a CA should properly determine the update period (U), considering both the efficiency and the guarantee range of business continuity.

5 Conclusion

The PKI is a widely used technique around the world change in various fields as a means of identification certification on networks. The role of CAs carrying on an important role of PKI is increasing in importance. In the event of a disaster in which the CA secret key is exposed for any reason, it is quite obvious that it would cause enormous monetary loss and many disruptions. Nevertheless, the research on this issue has been insufficient and few works that fully address this issue have yet been proposed. The existing PKI model [8] has some limitations such as inefficiency and impracticalness because the existing PKI models do not utilize the current PKI structure.

To the best of our knowledge, the disaster coverable PKI model proposed in this paper is the first model that maintains the existing PKI structure and ensures business continuity in the event of a CA's secret key exposure.

Acknowledgement

We would like to thank Yeon Hee Kim for sticking by us to complete this paper even though her life went in a different direction.

References

1. M. Abdalla, L. Reyzin, A New Forward-Secure Digital Signature Scheme, *Advanced in Cryptology - Asiacrypt 2000*, Lecture Notes in Computer Science, vol. 1976, pp. 116-129, Springer-Verlag, 2000.
2. R. Anderson, Invited lecture, *Fourth Annual Conference on Computer and Communications Security*, ACM, 1997.
3. M. Bellare, Sara K. Miner, A Forward-Secure Digital Signature Scheme, *In: CRYPTO'99*, LNCS 1666, pp. 431-448. Springer-Verlag, 1999.
4. Satoshi KOGA and Kouichi SAKURAI, Decentralization Methods of Certification Authority Using the Digital Signature Schemes, *2nd Annual PKI Research Workshop-Pre-proceedings*, 2003.

5. H. Krawczyk, Simple Forward-Secure Signature's From Any Signature Scheme, *In 7th ACN Conference on Computer and Communications Security*, 2000.
6. Zhengyi Le, Yi Ouyang, James Ford, Fillia Makedon. A Hierarchical Key-Insulated Signature Scheme in the CA Trust Model. *In: Information Security (ISC 2004)*, LNCS 3225, pp. 280-291. Springer-Verlag, 2004.
7. T. Malkin, D. Micciancio, and S. Miner, Efficient Generic Forward-Secure Signatures With An Unbounded Number Of Time Periods, *In Advances in Cryptology - EUROCRYPTO'02*, Lecture Notes in Computer science. Springer-Verlag. Vol. 2332, 2002.
8. Vesselin Tzvetkov, Disaster coverable PKI model based on Majority Trust principle, *Proceedings of the international Conference on Information Technology: Coding and Computing*, 2004.

A Corporate Capital Protection and Assurance Model

Colette Reekie and Basie von Solms

Academy for Information Technology
University of Johannesburg
APK Campus
P.O. Box 524
Auckland Park
Johannesburg, 2006
South Africa

cmt@rau.ac.za, basie@rau.ac.za

Abstract. This paper introduces the concept of Corporate Capital Protection Assurance. The authors provide a holistic Corporate Capital Protection Assurance model consisting of effective due diligence controls so that any organization regardless of its size or state of maturity can provide assurance to its members and stakeholders that all relevant 'Corporate Capital' (in the widest sense including aspects like intellectual capital, brand name, electronic assets, public opinion, trust, human capital, competitiveness etc) will be adequately protected. Corporate Capital Protection Assurance is more than information security protection of the confidentiality, integrity and availability of information. It includes the aspects mentioned above, as well as the policies, procedures, processes and human skills that must be protected. Therefore the authors have defined Corporate Capital Protection Assurance as the management commitment and leadership, with all the supporting people and structures all working together to provide for the adequate protection of the company's Corporate Capital. Thus Corporate Capital Protection Assurance entails more than information security and information security governance. It includes for e.g. the protection of a digital forensic infrastructure, aspects relating to risk management, to business continuity planning and control, to the protection of human resource information, knowledge and human resource skills, as well as the protection of information relating to policy formulation and content. All of these aspects need to be controlled in a formalized cohesive manner so that they are aligned with the overall business strategy and culture of the organization. This model will provide a consolidated view of all these above-mentioned types of corporate capital resources that cannot alone be protected by Information Security Governance controls and yet still require that require protection. Therefore this paper will provide a consolidated view of all these types of protection that should be provided by an organization, as well as provide a detailed exposition on the creation of and use of this Corporate Capital Protection Assurance model for organizations globally.

Keywords: Corporate Capital Protection and Assurance, Information Security Governance, IT Governance, Digital Forensic Governance, Corporate Governance, Project Management Governance.

1 Introduction

In light of the global move to ensure corporate governance objectives are met, senior management and boards of directors have realized that there is a greater need to account for the requirements of multiple stakeholders. These requirements have been the main driving force behind corporate governance. Corporate governance has thus begun to play an increasing role for all organizations globally. However, not only has corporate governance impacted the way in which organizations conduct business, more specifically its incumbent sub-governance domains such as Risk Governance, IT governance, Information Security Governance and Digital Forensic Governance now play an increasing role for all organizations globally.

The result of this is that it is no longer the sole responsibility of the CIO's and IT managers to adhere to these due diligence controls, but it has also caused the senior managers and boards of directors of organizations to become more aware of their information management responsibilities. As organizations develop and they become increasingly aware of their responsibilities for information resources, they need to ensure that their strategy for the protection of valuable information is continually evolving.

Research has shown that there is a need for a model that will guide organizations in their undertaking of compliance for IT governance, risk governance, information security governance and digital forensic governance. [10] [7] [11] More specifically this model needs to assist organizations in piecing all aspects of corporate governance together in a comprehensive format. The resulting model will therefore help guide organisations in the appropriate protection of their information resources. A clear understanding of an organizations position with regards to the implementation of IT governance, its attendant governance compliance areas as well as its sub compliance areas, such as information security governance, business continuity planning and digital forensic governance needs to be developed.

This paper proposes a corporate capital protection assurance model that will allow for:

- A clearer understanding of the heightened responsibilities of management with regard to information protection the reason for a differentiation to be made to that of information security.
- The need for a comprehensive corporate capital protection assurance model
- The creation of a visual depiction of the corporate capital protection assurance model that includes information security protection aspects from several governance domains.

2 Corporate Capital Protection Assurance

Before a model for corporate capital protection assurance can be established it is first necessary to define corporate capital protection assurance. This section will explain important concepts that play an important role in the defining of corporate capital protection assurance and will explain why information security alone is not enough to provide complete corporate capital protection.

Corporate Capital must be considered in the widest sense. This includes aspects such as intellectual capital, brand name, electronic assets, public opinion, trust, human capital, competitiveness i.e. corporate capital is so much more than information alone. The question now is, where do organizations look to start with their responsibility for due diligence?

This is where corporate governance comes into effect. The Cadbury report defines *Corporate governance* as "the system by which organizations are directed and controlled. Boards of directors are responsible for the governance of their organization." [1] Furthermore the OECD elaborates that "Corporate governance involves a set of relationships between a company's management, its board, its shareholders, and other stakeholders. Corporate governance also provides the structure through which the objectives of the company are set, and the means of attaining those objectives and monitoring performance are determined. Good corporate governance should provide proper incentives for the board and management to pursue objectives that are in the interests of the company and its shareholders and should facilitate effective monitoring." [2]

Despite CEO's and boards of directors looking after the organisation's objectives and stakeholders needs, a key challenge for today's knowledge based economy lies the increasing dependency on the derivation of value from IT systems. Adequate control and due diligence is expected by shareholders and key members, to protect valuable knowledge resources. Therefore IT governance, information security governance and digital forensic governance have become a business imperative.

IT Governance is defined by the IT Governance institute as "an integral part of enterprise (corporate) governance and consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extend the organisation's strategies and objectives." [4] Furthermore it is necessary to state that according to Gartner "IT governance specifies the decision-making authority and accountability to encourage desirable behaviours in the use of IT. IT governance provides a framework in which the decisions made about IT issues are aligned with the overall business strategy and culture of the enterprise. Governance is about decision making per se - not about how the actions resulting from the decisions are executed. Governance is concerned with setting directions, establishing standards and principles, and prioritizing investments; management is concerned with execution." [6]

Knowing the importance of IT governance, equally important subcomponents of IT governance have been identified as information security governance, human resource governance and business continuity planning.

Information Security governance consists of "the management commitment and leadership, organizational structures, user awareness and commitment, policies, procedures, processes, technologies and compliance enforcement mechanisms, all working together to ensure that the confidentiality, integrity and availability of the company's electronic assets (data, information, software, hardware, people etc) are maintained at all times. [7]

Human Resource governance is a relatively new organisational practice, and there is not as yet a commonly acknowledged definition. Sussman defines HR Governance as "the act of leading the HR function and managing related investments to optimise human capital performance, define stakeholders and their expectations, to fulfil fiduciary and financial responsibilities, to mitigate enterprise HR risk, and to assist

HR executive decision making.” [8] For the purposes of this research the authors felt it necessary to include HR governance as a component of IT governance because within any industry many breaches of security occur due to the human element. [12] It includes the policies procedures and awareness programs that need to be in place for the correct education of members of an organisation to allow for the adequate protection of their corporate capital.

Another key aspect in IT governance is the creation of a business continuity plan. The *business continuity plan* of an organization is “an all encompassing term covering both disaster recovery planning and business resumption planning.” [9][18] In other words in order to ensure the continuity of your business, it is necessary to have a plan in place that will allow your business to recover. This includes your IT systems to press relations and to the human resources required.

One more aspect of Corporate and IT governance that is newly emerging as a critical component is digital forensic governance. *Digital Forensic governance* has been defined as “the management commitment and leadership, organisational structures, procedures, processes and technologies all working together to ensure a proper environment for digital forensics to operate in.” [7]

The final three governance aspects relate to IT governance yet cannot be included within it per say as there are aspects relative to each that fall outside the scope of IT governance. These governance areas include, but are not limited to project governance, risk governance and policy governance.

Project governance is the term used in industry (especially within the IT sector) that describes “the processes that need to exist for a successful project that will outline the relationships between all internal and external groups involved in the project; that will also describe the proper flow of information regarding the project to all stakeholders and will ensure the appropriate review of issues encountered within each project including ensuring that required approvals and direction for the project is obtained at each appropriate stage of the project.” [3] “Project Governance extends the principle of IT Governance into the management of individual projects. Today, many organisations are developing ‘Project Governance Structures’. A Project Governance structure is different to an Organisation Structure in that it defines accountabilities and responsibilities for strategic decision-making per project. This can be particularly useful to project management processes such as change control and strategic (project) decision-making.” [5]

Risk governance and IT governance are often two terms that are interchanged frequently in business. However a formal approach to risk governance according to Charette is “that risk governance is integral to a corporation's complete process of governance. An assumption of good governance practice is that an effective risk management process exists that can ensure that the plethora of corporate compliance risks is addressed.” [13]

Finally another important component of corporate governance is policy governance. Carter has defined *Policy Governance*® “as a style of leadership that helps to define roles and relationships in which the Board functions as the visionary leader of a company, no longer involved in the daily operations.” [14] In other words the Board must therefore manage the affairs of the organization and cannot merely delegate with no follow up thereafter.

The next section provides a detailed exposition of the corporate capital protection assurance model that places all of these aspects above into perspective and goes on to explain how the model can help organisations implement a comprehensive and holistic approach to the protection of their corporate capital.

3 The Corporate Capital Protection Assurance Model

Based on the above definitions the authors have defined Corporate Capital Protection Assurance as “as the management commitment and leadership, with all the supporting people and structures all working together to provide for the adequate protection of the company’s Corporate Capital, including aspects such as intellectual capital, brand name, electronic assets, public opinion, trust, human capital and competitiveness. Figure 1 provides a visualization of the corporate capital protection assurance model.

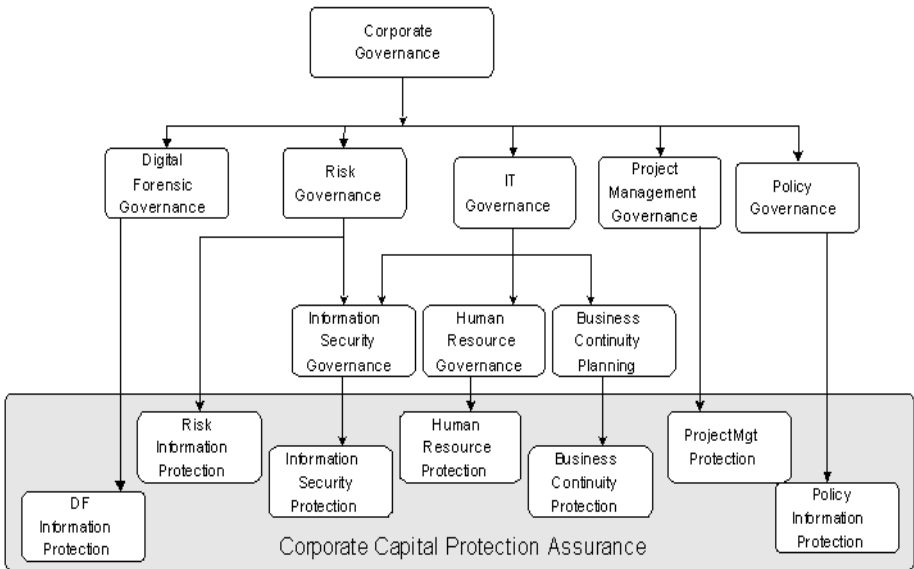


Fig. 1. A corporate capital protection assurance model

The model above provides a visualization of the corporate capital protection assurance model and illustrates the components of various organisational and governance aspects. From the above diagram it is evident that there are matters relating to corporate capital resources that need to be protected and yet cannot rely on protection from information security alone. In this initial research, the authors identified several governance domains that can assist in the protection of corporate capital resources. They have been identified as digital forensic governance, risk governance, IT governance, project management governance and policy governance. Sub-domains of IT Governance were identified as information security governance, human resource governance and business continuity planning. The following section

will describe terminology that has been illustrated in the Corporate Capital Protection Assurance model (as depicted in the greyed area):

3.1 Digital Forensic Information Protection

Crucial to the success of any digital forensic process is proving the integrity of the evidence as well as the integrity of the investigation process. [16] [17][15] Aspects of the digital forensic investigation process fall outside of the scope of Information Security Governance and hence relate to information that has to be specifically protected. This information includes information relating to the gathering of evidence and the ethical usage of this information. This is the responsibility of the digital forensic team to ensure. In light of this organisations have focused attention on accountability and hence good due diligence over the digital forensic process. In some instances this digital forensic information requires that the fiduciary, effectiveness and efficiency requirements are also protected by the investigators.

3.2 Risk Information Protection

Risk governance can be viewed from two perspectives, namely from an information security control perspective and from a risk assessment perspective. If for example an IT software implementation could incur excess costs, the issue becomes a risk governance protection issue because it could impact on the financial risk of the organisation. This could ultimately impact the reputation of the organisation. Therefore certain information relating to the protection of risk information that falls outside of the scope of information security and IT governance must be monitored.

3.3 Information Security Protection

Information security protection is the area relating to the protection of the processes, policies and procedures for the protection of information. It is more than the mere data that needs protecting. It includes the information security skills and knowledge that the organisation has at its disposal that must also be protected.

3.4 Human Resource Information Protection

Human resource information protection is more than information security protection because it includes more than the protection of data relating to its employees. It includes the protection of the processes for employing human resources, training of these human resources and ultimately the protection of the organisation's time and effort spent on development of its employees.

3.5 Business Continuity Information Protection

As stated earlier it is necessary to ensure the ongoing operation of the organisation including the IT systems to the human resources required and the management of press relations in the event of a disaster. [18] Business continuity information protection will allow for the protection of information and processes relating to the critical business processes of an organisation.

3.6 Project Management Information Protection

Project management information protection includes the defining of responsibilities and accountability of key role players in the development of a project. Furthermore this implies that it is more than the data related to an IT project, it can also be the protection of any project that falls in the scope of the organisations strategic project portfolio.

3.7 Policy Information Protection

This entails the satisfactory protection of human resources, process and information at a board level that needs timely information to make effective decisions.

It is evident that the corporate capital protection assurance model allows for the specific defining of information related to information security as well as going beyond the basic IT governance and information security governance domains. The following section goes onto conclude and elaborate on future developments.

4 Conclusion and Future Developments

The corporate capital protection assurance model above illustrates an initial research model into the protection of more than information security data. It includes aspects such as the protection of policies, procedures, processes, brand reputation, and human skills. This model has resulted in a need being identified for the development of a more detailed maturity model. This maturity model will provide assistance to organizations in the development of a plan for the protection of their corporate capital resources, in accordance with their varying sizes and stages of development. Therefore it is evident that organizations need a model that can show them where to begin and how to progress through to a mature state of IT governance compliance whilst still keeping in line with the other areas of corporate capital protection assurance. This will be the topic for further research.

References

1. Cadbury, A. "The Cadbury Report." Institute of Internal Auditors. 1992. Available online: <http://www.iaa.org.uk/cms/IIA/uploads/2c9103-ea9f7e9fbe--7e3a/Cadbury.pdf>
2. Organisation for Economic Co-operation and Development. "OECD Principles of Corporate Governance: 2004." OECD, 2004. Available online: http://www.oecd.org/document/49/0,2340,en_2649_34813_31530865_1_1_1_1,00.html.
3. Answers.Com. "Project Governance." Available online: <http://www.answers.com/topic/governance>
4. Breeding Alex. "Make IT Governance an integral part of the enterprise." IT Governance Institute 2003. Available online: <http://www.itgi.org>
5. Project Management Informed Solutions. "Project Governance." Accessed on 7July 2006. http://www.pmis.co.uk/project_governance.htm
6. Dallas, Susan, Bell, Michael. "The Need for IT Governance: Now More Than Ever (AV-21-4823)." Gartner, 20 January 2004.

7. Von Solms, SH, Louwrens, CP. Chapter X “Digital Crime and Forensic Science in Cyberspace – the relationship between Digital Forensics, Corporate Governance, Information Technology Governance and Information Security Governance” Edited by P. Kanellis, N. Kolokotronis, E. Kiountouzis, D. Martakos Idea Group Inc. 2005.
8. Sussman, M. “Why HR Governance Matters.” July 2006. Available online: http://www.ceoforum.com.au/200406_remuneration.cfm
9. Calpoly. “Disaster scope” Available online: http://ccs.calpoly.edu/printable/disaster_scope.html
10. Van Grembergen, W. “Strategies for information technology governance.” Idea group publishing. United Kingdom. 2004.
11. Petersen, R. “Integration Strategies and Tactics for information technology governance.” Idea Group Inc. 2004.
12. Smith, R and Gordon, L. “2005 CSI/FBI computer crime and security survey” Computer Security Institute. 2005
13. Charette, RN. “Risk Governance Understood” CIO.com 2005. Available online: <http://www2.cio.com/analyst/report3298.html>
14. Carver, J. “The Carver model of Policy Governance” Policy governance model, 2006. Available online: <http://www.carvergovernance.com>
15. Sommer. 1998. “Intrusion Detection Systems as Evidence.” Available online. http://www.raid-symposium.org/raid98/ProgRAID98/Full_Papers/Sommertext.pdf
16. Stephenson. 2002. “End – to – End Digital Forensics”. Computer Fraud and Security. Volume 2002, Issue 9
17. Stephenson. 2003. “Using evidence effectively” Computer Fraud and Security. Volume 2003 Issue 3.
18. State Records Authority. “Business Continuity Planning based on AS/NZS 17799 Clause 11.1” New South Wales, 2003. Available online: <http://www.records.nsw.gov.au/publicsector/rk/glossary/singleversion.htm>

Quantitative Evaluation of Systems with Security Patterns Using a Fuzzy Approach

Spyros T. Halkidis, Alexander Chatzigeorgiou, and George Stephanides

Department of Applied Informatics, University of Macedonia
Egnatia 156, GR-54006, Thessaloniki, Greece
halkidis@java.uom.gr, {achat, steph}@uom.gr

Abstract. The importance of Software Security has been evident, since it has been shown that most attacks to software systems are based on vulnerabilities caused by software poorly designed and developed. Furthermore, it has been discovered that it is desirable to embed security already at design phase. Therefore, patterns aiming at enhancing the security of a software system, called security patterns, have been suggested. The main target of this paper is to propose a mathematical model, based on fuzzy set theory, in order to quantify the security characteristics of systems using security patterns. In order to achieve this we first determine experimentally to what extent specific security patterns enhance several security aspects of systems. To determine this, we have developed two systems, one without security patterns and one containing them and have experimentally determined the level of the higher robustness to attacks of the latter. The proposed mathematical model follows.

Keywords: Software Security, Security Patterns, Fuzzy Risk Analysis.

1 Introduction

The importance of software security has been evident since the discovery that most attacks to real software systems are initiated by software poorly designed and developed [34, 32, 15, 16]. Furthermore, it has been shown that the earlier we incorporate security in a software system the better [34]. Therefore, in analogy to design patterns [13], which aim at making software well structured and reusable, Security Patterns [33, 4] have been proposed, targeting at imposing some level of security to systems already at the design phase.

In this paper, we try to propose a mathematical model for the security of systems using security patterns. To achieve this, we first investigate to what extent specific security patterns reinforce several aspects of software systems security. To determine this experimentally we have built two software systems, which are the implementations of web applications, one without security patterns and one where security patterns were added to the former. We studied all applications under known categories of attacks to web applications [29]. To perform our analysis we have used the AppScan Web Application Penetration Testing tool, and organized a contest to study other approaches for evaluating software systems for vulnerabilities. We have estimated experimentally to what extent the system using security patterns is more

robust to attacks compared to the one that does not use them. Furthermore, initiated by the findings, we propose expressions for the resistance to STRIDE attacks [16] for the patterns examined. Finally, we use results from fuzzy reliability [7] and the application of fault trees [6, 1], to examine the security properties of systems using security patterns and illustrate the application of the related results to a system properly using the security patterns examined.

The remainder of the paper is organized as follows. In Section 2 we briefly review work on security patterns. Section 3 is a description of the systems under examination. In Section 4 we describe the results of our evaluation. Section 5 proposes a mathematical model for systems using security patterns using fuzzy numbers and fuzzy fault trees. Finally, in Section 6 we make some conclusions and propose future research directions.

2 Security Patterns

Since it has been evident that it is desirable to incorporate security already at the design level [34, 16], various efforts to propose security patterns, that serve this aim, have been done.

Yoder and Barcalow were the first to propose security patterns [35] in 1997. Since then, various security patterns were introduced. Patterns for enterprise applications [27], patterns for authentication and authorization [11, 20], patterns for web applications [18, 36], patterns for mobile java code [23], patterns for cryptographic software [5] and patterns for agent systems [24]. Though, all these efforts did not share some common terminology.

The first effort to provide a comprehensive review of existing security patterns was done by the OpenGroup Security Forum [4]. In this work, security patterns are divided into Available System Patterns, which are related to fault tolerance [25] and Protected System Patterns, which aim at protecting resources.

In an earlier work [14] we have performed a qualitative evaluation of these security patterns.

Recently, a summary of security patterns has appeared in the literature [33]. In this text security patterns are divided into web tier security patterns, business tier security patterns, security patterns for web services, security patterns for identity management and security patterns for service provisioning. In this paper we focus on web tier security patterns and business tier security patterns.

3 Description of the Systems Under Examination

In order to perform our security analysis, we have used two systems. Specifically, we have developed a simple e-commerce application without security patterns, hereafter denoted as “first” application, and a second application where security patterns were added to it, hereafter denoted as “second” application.

The first application under consideration is a typical J2EE (Java 2 Enterprise Edition, now referred to as Java EE) application with no security patterns. We have chosen J2EE as a platform for both applications since the J2EE platform is widely

used in business applications and is useful from the security point of view [33, 3]. In our systems we have used JBoss 4.0.3 as an application server that encompasses the web and business tier, and MySQL 5.0 for the database tier.

The first system consists of 46 classes. It has 16 servlets and 7 EJBs. One EJB works as a web service endpoint [26].

We have left on purpose on this system so-called “security holes” that attackers can exploit.

First of all, several sources for SQL injection [29, 2, 31, 12] were included. An SQL injection attack occurs when an attacker is able to insert a series of SQL statements into a query that is formed by an application, by manipulating data input that is not properly validated [2]. SQL injection attacks can cause unauthorized viewing of tables, database table modification or even deletion of database tables.

Furthermore, several sources for cross-site scripting were included. Cross site scripting [29, 10, 30, 17], also known as XSS, occurs in a web application when data input in one page which are not properly validated, are shown in another page. In this case, script code can be input in the former page that is consequently executed in the latter. In this way it is easy to perform an Information Disclosure attack [16] for example by using Javascript code that shows the cookie values of sensitive information.

Additionally, several sources for HTTP Response Splitting [19], were included in the application. HTTP Response Splitting attacks can occur when user data that were not properly validated are included in the redirection URL of a redirection response, or when data that were not properly validated are included in the cookie value or name, when the response sets a cookie. In these cases, by manipulating http headers, it is easy to create two responses instead of one where in the second response an XSS attack can be performed. Variants of this attack include Web Site Defacement, Cross User page defacement, Hijacking pages with user specific information and Browser Cache Poisoning [19].

Furthermore, in the first application no SSL connection was used and therefore sensitive information such as credentials and important information in cookies could be eavesdropped.

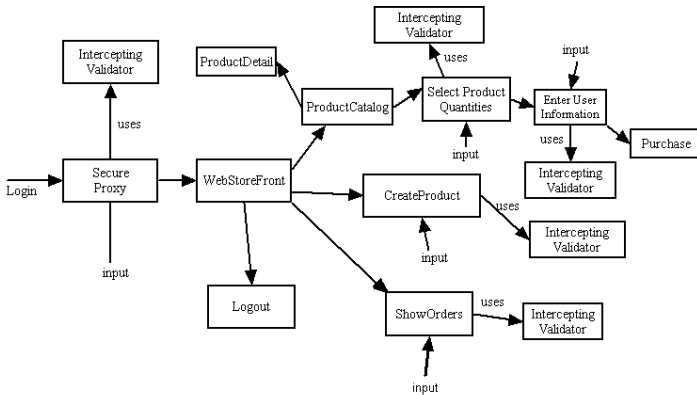


Fig. 1. Block diagram of the second application under examination

Finally, servlet member variables race conditions were included, which could be exploited by a number of users acting simultaneously.

In the second application we have built, the sources for attacks were not removed, but security patterns were used with the aim of protecting against them. The second application consists of 62 classes. It has 17 Servlets and 9 EJBs where one EJB again serves as an endpoint for the web service. The security patterns used in this system are the Secure Proxy pattern, Login Tunnel variant [4], the Secure Pipe pattern, the Secure Logger pattern, Secure Log Store Strategy, the Intercepting Validator pattern and the Container Managed Security pattern [33]. In Figure 1 we show a block diagram that consists of the main components of the second application with some of the security patterns used. Solid arrows show the flow of information.

4 Evaluation of the Systems with Regard to Attacks

In order to evaluate the systems with regard to attacks, we have used Watchfire's AppScan web application penetration testing tool. Furthermore, we have initiated a web application security contest, which was won by Benjamin Livshits from Stanford University. Livshits used static analysis tools to find the security flaws which are described in several papers [21, 22].

Both approaches found the major security flaws of the applications, meaning SQL Injection and Cross Site Scripting vulnerabilities. However both approaches had several false positives. AppScan for example found sources for buffer overflows, while java was used and the static analysis approach found sources for SQL injection in the second application, by examining the code for the EJBs, while proper input validation was done at the Web Tier. Race conditions for servlet member variables were found only by the static analysis approach. Several application errors of low severity not found by the static approach, were found by AppScan (checking for proper session variable values, that though not cause security risks). AppScan found the unencrypted login request flaw in the first application that did not use SSL. AppScan also found unencrypted SSL parameter flaws in the second application, which are of low severity. HTTP response splitting attacks in the first application as well as race conditions existing in the third application were found by neither of the approaches.

Additionally, the security flaws found by both approaches, were fewer in the case of the second application in comparison to the first one. The difference between the number of flaws found for the first and the second application was much more prominent in the set of high-risk flaws.

After careful analysis of the results we concluded that proper use of the security patterns leads to remediation of all the security flaws, except flaws that are of minor risk, like unencrypted SSL parameters (of course this flaw is of minor risk only when the unencrypted parameters are not crucial like in our case). These flaws that remain even after the use of security patterns, are due to the degrees of freedom left to the programmer even after using them imposes some level of security. Furthermore, current security patterns impose no rules for the use of servlet member variables and therefore race conditions may remain in a system using security patterns.

The Intercepting Validator pattern, when used for all input, including session variables, and variables that are not input by the user but still posted, protects from SQL Injection, Cross-Site scripting, and HTTP Response Splitting attacks. It offers therefore very high resistance to Tampering with Data and Information Disclosure Attacks [16].

The Secure Proxy pattern, Login Tunnel variant, has two levels of authentication in order to protect from Spoofing Identity, Elevation of Privilege and Information Disclosure attacks. Its resistance to related attacks can be estimated by considering it to be the equivalent the protection of two guards [4] connected in a series. The resistance of both of these patterns to attacks is dependent to the robustness of the authentication mechanism to dictionary attacks. Recent studies [37, 28] have shown that dictionary attacks, with a usual distribution of the complexity of the passwords selected, succeed 15-20% of the times. The authentication mechanism of the Protected System pattern can still be marked as of high security. All authentication patterns and consequently these two patterns examined here should be resistant to eavesdropping attacks to serve their purpose. Therefore, they should always be used in combination with the Secure Pipe pattern that provides SSL encryption.

The Secure Pipe pattern offers protection from information disclosure attacks. The programmer can still use unencrypted parameters in an SSL request, but usually, when these parameters are not of crucial importance this kind of flaw is of minor risk.

The Container Managed Security Pattern implements an authorization mechanism. It protects from Elevation of Privilege, Information Disclosure and partly from Spoofing Identity attacks, since anyone who belongs to the Role allowed to access the EJBs could do so.

Table 1. Resistance of the security patterns examined to STRIDE attacks

	S	T	R	I	D	E
Intercepting Validator		very high		very high		
Guard of Secure Proxy with Secure Pipe	high			high		high
Container Managed Security	medium			very high		very high
Secure Logger		very high				

Finally, the Secure Logger pattern protects from tampering the log created.

The evaluation of these security patterns with respect to the STRIDE (Spoofing Identity, Tampering with Data, Repudiation, Information Disclosure, Elevation of Privilege) model [16] is summarized in Table 1. The irrelevant entries are left blank.

5 Fuzzy Mathematical Model for Systems Using Security Patterns

One of the targets in our research was to build a mathematical model for systems that use security patterns, based on our findings for the level of security each pattern

offers. The most appropriate models for our purpose seem to be risk analysis models [1]. We have chosen to use a fuzzy risk analysis model because it is impossible to determine security characteristics of software systems using exact numbers. As Høglund and McGraw [15] note, in software risk analysis exact numbers as parameters work worse than having values such as high, medium and low. These kinds of values can be termed as fuzzy.

Risk analysis techniques for estimating the security of systems have been proposed earlier [1]. The differences in our approach are that we apply risk analysis already at the design phase of a software system using a security pattern centric approach, that we make use of the newer STRIDE model of attacks [16] and that we use fuzzy terms.

When performing risk analysis for a system, a common formula used by the risk engineering community is the following [8]:

$$R = LEC . \tag{1}$$

where *L* is the likelihood of occurrence of a risky event, *E* the exposure of the system to the event, *C* the consequence of the event and *R* the computed risk. Examining this equation in comparison to the risk analysis performed by Høglund and McGraw [16] in our case the likelihood *L* is the likelihood of a successful attack, the exposure *E* is a measure of how easy is to carry out the attack and *C* is the impact of the attack. As we explained earlier we have chosen that the terms in our risk analysis model are fuzzy.

Table 2. Mapping of linguistic terms to generalized fuzzy numbers

Linguistic Term	Generalized Fuzzy Number
absolutely-low	(0.0, 0.0, 0.0, 0.0; 1.0)
very-low	(0.0, 0.0, 0.02, 0.07; 1.0)
low	(0.04, 0.1, 0.18, 0.23; 1.0)
fairly-low	(0.17, 0.22, 0.36, 0.42; 1.0)
medium	(0.32, 0.41, 0.58, 0.65; 1.0)
fairly-high	(0.58, 0.63, 0.80, 0.86; 1.0)
high	(0.72, 0.78, 0.92, 0.97; 1.0)
very-high	(0.93, 0.98, 1.0, 1.0; 1.0)
absolutely-high	(1.0, 1.0, 1.0, 1.0; 1.0)

The applicability of fuzzy techniques to security problems has already been proposed [8] and the use of fault trees for security system design has also been suggested [6, 1]. In this paper we perform an analysis of the security of systems using security patterns, using results from fuzzy set theory [38] and fuzzy fault trees [7]. Specifically, we perform fuzzy risk analysis for a system that has properly added security patterns to the initial system under examination.

Our analysis uses generalized fuzzy numbers [9] and the similarity metric proposed by Chen and Chen [9]. We have chosen generalized fuzzy numbers instead of other existing approaches because the similarity measure for generalized fuzzy numbers has been proven to be robust in the cases where both crisp and fuzzy numbers are to be compared [9].

We used the mapping from linguistic terms to generalized fuzzy numbers shown in Table 2 adapted from [9]:

These fuzzy numbers (except absolutely-low and absolutely-high) are shown in Figure 2.

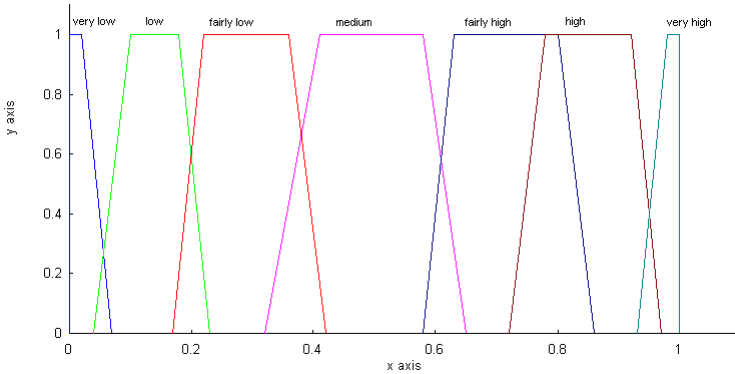


Fig. 2. Fuzzy numbers that correspond to the linguistic terms used in our analysis

Table 3. Analysis of primary attack events for a system properly using security patterns

Primary Event	Likelihood of occurrence	Exposure	Consequences	Categories of Attacks
Event 1. Dictionary attack to a guard of Secure Proxy is successful	low	high	very high	S, E, I
Event 2. Variable value is used unencrypted in SSL request	high	very high	low	I
Event 3. Variable value is read from wsdl file	medium	very high	low	I
Event 4. Input validation is bypassed	absolutely low	high	very high	T, I
Event 5. Unauthorized access to servlet member variables is allowed by exploiting race conditions	high	low	low	I

We then identified the primary events for the fault trees and the categories of attacks related to the STRIDE model [16] they belong to. A dictionary attack to the Secure Proxy pattern is successful only if both guards are compromised and causes a Spoofing Identity, Elevation of Privilege and Information Disclosure. An attack to a guard of this pattern can be performed using automated tools and therefore the exposure for this attack is high. The likelihood of such attack is low since the guard has high resistance to dictionary attacks. If such an attack is successful the consequences are very high. By performing a similar likelihood-exposure-consequence analysis for all primary events we obtain Table 3.

The Tampering with data attack does not exist practically for this system, since the only primary event that causes it has absolutely low likelihood of occurrence. The Spoofing Identity and Elevation of Privilege attacks occur for the same primary event.

The resulting fuzzy fault tree for Information Disclosure attacks is shown in Figure 3. The fault tree for Spoofing Identity and Elevation of Privilege attacks can be built using the same technique.

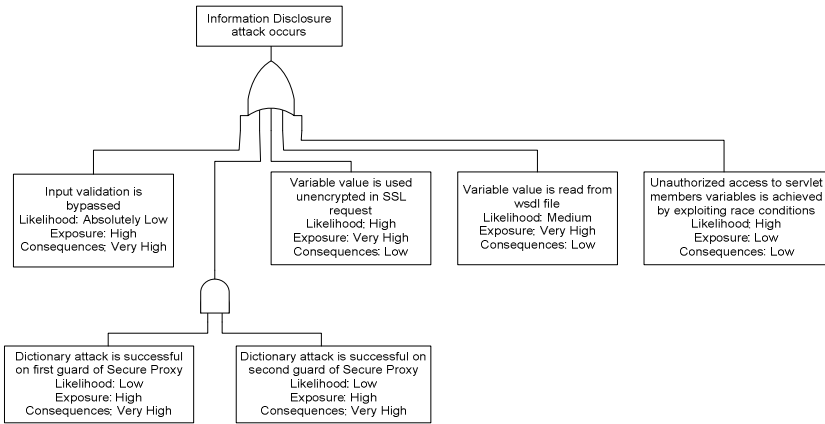


Fig. 3. Fault tree for Information Disclosure attacks

The methodology we use to derive the risk for the top event is outlined in the following steps:

- 1) We first identify the values of likelihood, exposure and consequences for the primary events.
- 2) We then perform the logical composition of values, according to rules for the gates of fault trees, starting from the values of primary events and ending at the computation of the risk for top event.
- 3) Finally we compare the risk for the top event computed in step 2, with the values in Table 2 using the similarity metric from [9].
- 4) The linguistic term with the highest similarity is chosen as the result.

This is a typical fuzzy risk analysis approach [7] where the terms for the events depend though on the security patterns used in the system examined. Furthermore, we use in our analysis generalized fuzzy numbers adapted from [9] as well as the similarity metric from [9].

Table 4. Summary of risks computed for different types of attacks for a system without security patterns and a system properly using them

	Spoofing	Tampering with Data	Information Disclosure	Elevation of Privilege
System without security patterns	fairly high	fairly high	high	fairly high
System properly using security patterns	very low	absolutely low	low	very low

After performing the necessary computations for the system properly using security patterns we come to the result that the risk for the fault tree corresponding to Spoofing and Elevation of Privilege attacks is *very low* and the risk for the fault tree corresponding to Information Disclosure attacks is *low*. These trees correspond to the system that properly uses security patterns. The risk for Tampering with data attacks is zero (*absolutely low*).

On the contrary, for the system that does not employ security patterns, the risk values according to the proposed model, for the same types of attacks are *fairly high* for Spoofing Identity and Elevation of Privilege attacks, *high* for Information Disclosure attacks and *fairly high* for Tampering with data attacks. Table 4 summarizes these results and quantifies the difference between the two systems.

The methodology described thus allows us to derive results about the *total* security of systems employing security patterns, already at the design, in terms of fuzzy linguistic variables.

7 Conclusions and Future Work

The results of the evaluation of the attacks as well as the fuzzy methodology used show that systems that use security patterns properly are highly secure and robust to attacks. This robustness to attacks has been also quantified in this work and a mathematical model has been proposed. Future work includes the introduction of new security patterns that solve the issues not addressed by existing ones and a software tool that automates the security evaluation process we described in this paper.

Acknowledgements

We would like to thank the Web Application Security mailing list of SecurityFocus and the comp.lang.java.security mailing list, for letting us organize a contest. Furthermore, we would like to thank Benjamin Livshits, from Stanford University, the winner of the contest and Watchfire Corporation for providing us an evaluation license for AppScan.

References

1. Amoroso, E., Fundamentals of Computer Security Technology, Prentice Hall (1994)
2. Anley, C., Advanced SQL Injection in SQL Server Applications, NGSSoftware whitepaper (2002)
3. Berry, C. A., Carnell, J., Juric, M.B., Kunnumpurath, M. M., Nashi, N. and Romanosky, S., J2EE Design Patterns Applied, Wrox Press (2002)
4. Blakley, B., Heath, C. and Members of the Open Group Security Forum, Security Design Patterns, Open Group Technical Guide (2004)
5. Braga, A., Rubira, C., and Dahab R., Tropyc: A Pattern Language for Cryptographic Software, in Proceedings of the 5th Conference on Pattern Languages of Programming (PLoP '98) (1998)
6. Brooke, P. J., and Paige, R. F., Fault Trees for Security System Design and Analysis, Computers and Security, vol. 22, No. 3, pp. 256-264 (2003)

7. Cai, K.-Y., *Introduction to Fuzzy Reliability*, Kluwer Academic Publishers (1996)
8. Cai, K.-Y., *System Failure Engineering and Fuzzy Methodology, An Introductory Overview*, *Fuzzy Sets and Systems*, Vol. 83 pp. 113-133 (1996)
9. Chen, S.-J., and Chen, S.-M., *Fuzzy Risk Analysis Based on Similarity Measures of Generalized Fuzzy Numbers*, *IEEE Transactions on Fuzzy Sets and Systems*, Vol. 11, No. 1 (2003)
10. Cgisecurity.com, *Cross Site Scripting questions and answers*, <http://www.cgisecurity.com/articles/xss-faq.shtml>
11. Fernandez E., *Metadata and authorization patterns*, <http://www.cse.fau.edu/~ed/MetadataPatterns.pdf> (2000)
12. Friedl, S., *SQL Injection Attacks by Example*, <http://www.unixwiz.net/techtips/sql-injection.html>
13. Gamma, E., Helm, R., Johnson, R., and Vlissides, J., *Design Patterns, Elements of Reusable Object-Oriented Software*, Addison Wesley (1995)
14. Halkidis, S. T., Chatzigeorgiou, A., and Stephanides, G., *A Qualitative Evaluation of Security Patterns*, in *Proceedings of the 6th International Conference on Information and Communications Security (ICICS '04)* (2004)
15. Hoglund, G. and McGraw, G., *Exploiting Software, How to Break Code*, Addison Wesley (2004)
16. Howard, M. and LeBlanc, D., *Writing Secure Code*, Microsoft Press (2002)
17. Hu, D., *Preventing Cross-Site Scripting Vulnerability*, SANS Institute whitepaper (2004).
18. Kienzle, D., and Elder, M., *Security Patterns for Web Application Development*, Univ. of Virginia Technical Report (2002)
19. Klein, A., "Divide and Conquer", *HTTP Response Splitting, Web Cache Poisoning Attacks and Related Topics*, Sanctum whitepaper (2004)
20. Lee Brown, F., Di Vietri J., Diaz de Villegas G., Fernandez, E., *The Authenticator Pattern*, in *Proceedings of the 6th Conference on Pattern Languages of Programming (PLOP '99)* (1999)
21. Livshits B., and Lam, M. S., In *Proceedings of the 14th USENIX Security Symposium* (2005)
22. Livshits, B., and Lam, M. S., *Finding Security Vulnerabilities in Java Applications with Static Analysis*, Stanford University Technical Report (2005)
23. Mahmoud, Q., *Security Policy: A Design Pattern for Mobile Java Code*, in *Proceedings of the 7th Conference on Pattern Languages of Programming (PLOP '00)* (2000)
24. Mouratidis, H., Giorgini, P., and Schumacher, M., *Security Patterns for Agent Systems*, in *Proceedings of the Eighth European Conference on Pattern Languages of Programs (EuroPLOP '03)* (2003)
25. Pullum, L. L., *Software Fault Tolerance Techniques and Implementation*, Artech House Publishers (2001)
26. Roman, E., Sriganesh, R. P. and Brose G., *Mastering Enterprise JavaBeans, Third Edition*, Wiley Publishing (2005)
27. Romanosky, S., *Enterprise Security Patterns*, <http://www.romanosky.net/papers/EnterpriseSecurityPatterns.pdf> (2002)
28. Ross, B., Jackson, C., Miyake, N., Boneh, D., and Mitchell, J.C., *Stronger Password Authentication Using Browser Extensions*, In *Proceedings of the 14th USENIX Security Symposium* (2005)
29. Scambray, J., and Shema, M., *Hacking Exposed Web Applications*, McGraw-Hill (2002)
30. Spett, K., *Cross-Site Scripting, Are your web applications vulnerable?*, SPI Labs whitepaper

31. SPI Labs, SQL Injection, Are Your Web Applications Vulnerable?, SPI Labs whitepaper.
32. Spinnelis, D., Code Quality : The Open Source Perspective, Addison Wesley (2006)
33. Steel, C., Nagappan R., and Lai, R., Core Security Patterns, Best Practices and Strategies for J2EE, Web Services, and Identity Management, Prentice Hall (2006)
34. Viega, J., and McGraw, G., Building Secure Software, How to Avoid Security Problems the Right Way, Addison Wesley (2002)
35. Yoder, J., and Barcalow, J., Architectural Patterns for enabling application security, in Proceedings of the 4th Conference on Pattern Languages of Programming (PLoP '97) (1997)
36. Weiss, M., Patterns for Web Applications, in Proceedings of the 10th Conference on Pattern Languages of Programming (PLoP '03) (2003)
37. Wu, T., A Real-World Analysis of Kerberos Password Security, In Proceedings of the 1999 Network and Distributed System Symposium (1999)
38. Zimmerman, H.-J., Fuzzy Set Theory and its Applications, Third Edition, Kluwer Academic Publishers (1996)

Managing Critical Information Infrastructure Security Compliance: A Standard Based Approach Using ISO/IEC 17799 and 27001

Wipul Jayawickrama

Infoshield Consulting
wipul@infoshield.com.au

Abstract. Information technology constitutes a substantial component of the critical infrastructure of many nations. Systems used by utilities and service industries such as electricity, water, wastewater treatment and gas are key components of these critical infrastructures. These critical infrastructures rely on a range of technologies commonly known as Process Control Systems in the production, distribution or management aspects of their services.

To ensure continued delivery of these critical services, it is important to ensure that the process control systems used to control, monitor and manage the infrastructure are secured against physical and cyber security threats. A number of information security standards have been defined by various industry and government regulatory bodies to provide guidance in securing process control systems. However, managing compliance to several standards can become an added administrative overhead to organizations.

This paper reviews the challenges in maintaining compliance with multiple standards and postulates that a holistic information security management system is required to ensure ongoing security of these process control systems. It proposes the implementation of international standards ISO/IEC 17799 and 27001 as a practical approach to managing the various compliance requirements and providing a framework to implement, monitor, manage and improve the security of process control systems.

1 Introduction

Developed nations and many developing nations rely heavily on Information Technology infrastructures to sustain and support the modern way of life. This reliance on computer systems and networks that constitute these infrastructures pervade across diverse activities such as supporting the ‘information economy’, maintaining national security and controlling the basics of every day life such as access to clean water. Any disruptions to the functions of these infrastructures can result in the disruptions to the way of life of the population.

Proliferating cyber security threats to computer networks and systems have prompted many governments to implement strategies to identify and safeguard those systems and networks that are critical to the nation. These strategies generally are developed and implemented under a ‘Critical Infrastructure Protection’ (CIP) program.

The Commonwealth Government of Australia defines critical infrastructure as "... those physical facilities, supply chains, information technologies and communication networks which, if destroyed, degraded or rendered unavailable for an extended period, would significantly impact on the social or economic well-being of the nation, or affect Australia's ability to conduct national defence and ensure national security [1]".

This definition of critical infrastructure is not limited to 'physical' elements of the infrastructure alone. It encompasses social, political and economical aspects of everyday life of the nation. As stated by the Attorney General's Department in a publicly released discussion paper in 2004, "banking and finance, transport and distribution, energy utilities, health, food supply and communications as well as key government services and national Icons", and the complex network of producers, processors, all contribute to the Australian way of life and thus are considered part of the critical infrastructure for the purposes of national critical infrastructure protection strategy [2].

A key component of the critical information infrastructure is Process Control Systems (PCS) used to monitor and control processes in various industrial environments. Commonly known as Supervisory Control and Data Acquisition (SCADA) systems, these control systems are used in critical utilities and service industries. Energy production and distribution, water and waste water treatment and water supply and distribution are some examples of critical infrastructures where these PCS and SCADA Systems are used.

Historical evidence shows that an information security breach of a process control system can cause damaging consequences. The 2000 Maroochydhore incident where a disgruntled ex-employee hacked in to a process control system and released sewage in to waterways [3], [4], the August 2003 US power grid blackout supposedly caused by the Blaster worm [5], and the January 2003 blaster worm which disabled safety monitoring systems at a Ohio-based nuclear power plant for nearly five hours [6] demonstrate the negative impacts of such a security breach.

This paper reviews the security problems associated with process control systems and argues that a holistic approach needs to be taken in securing the systems that are used to monitor and control critical information infrastructure. It proposes the use of international standard ISO/IEC 17799 as a viable approach to securing process control systems.

2 Defining Process Control Systems

A process control system is not a single device or a computer. It is a multi-tiered network of computers, logic controllers, human-machine interfaces, process control equipment and software applications. These systems allows an operator in a central control room to remotely interact with a distributed process; for example "... to make point changes to distant process controllers, to open or close valves or switches, to monitor alarms, and to gather measurement information Typically process control systems are applied to processes that span wide geographical areas". They are used to control and monitor processes that require "frequent, regular or immediate intervention" [7] by human operators who investigate and respond to these alarm conditions [8].

Following are examples of the application of process control systems [7]:

- hydroelectric generating stations – to control the turning on and off of valves to the turbines in response to customer demands on the power grid
- oil or gas production facilities – to control fluid measurement equipment and pumps, turning motors on and off, gathering meter information regularly, need to respond quickly to varying pressure conditions in the pumps and pipes
- pipelines for oil, gas, chemicals and water – to control the opening and closing valves, starting and stopping pumps, and responding quickly to leaks of dangerous or environmentally sensitive materials
- electric transmission systems – to control the opening and closing switches and responding quickly to load changes on the lines
- irrigation systems – to control the opening and closing valves and gathering meter values for the water supplied to consumers

Many of these uses typically are in environments that require 24 hour availability, near real-time response capability and the ability to control and respond remotely.

The term Process Control Systems is a generic term applied to several types of systems. These systems include Supervisory Control and Data Acquisition (SCADA) systems, Distributed Control Systems (DCS), Energy Management Systems (EMS), Safety Instrumented Systems (SIS), and Manufacturing and Control systems (M&CS) [9].

3 Safeguarding Process Control Systems – Challenges

There has been an increase in information security incidents related to control systems in the recent years. Analysis of incident reports recorded in the Industrial Security Incident Database (ISID) maintained by the British Columbia Institute of Technology (BCIT) indicate that there has been a five fold increase in the annual control system incident rate [9][10]. While petroleum, transportation, power and utilities industries were the objects of most of these incidents, chemical pulp and paper, water and waste water, electronic manufacturing, food and beverage, aerospace, and metals industries have reported process control system security incidents in the recent years [9].

Increasing incident rates pose a challenge to safeguarding the national critical infrastructure as nations are dependent on these industries to a very large extent. Miller identifies 14 areas that US National Strategy for Homeland Security has identified as critical infrastructure. As Miller states, these systems “support our everyday lives, from the water and food in our homes to our physical and financial welfare. They also support government and operation”. According to Miller “nearly every one of these critical infrastructures” uses process control systems [8].

One of the challenges to protecting the process control systems that constitute the critical information infrastructure systems is that not all of these systems are government owned. Of the 14 defined areas of critical infrastructure in the US, most are privately owned [8]. As the Attorney General’s Department of the Australian Commonwealth Government identifies as much as 90 per cent of the critical infrastructure in Australia is privately owned, which makes it impossible for the government alone to carry out protective measures. This calls for coordination and

participation of "... owners and operators of infrastructure, regulators, professional bodies and industry associations, in cooperation with all levels of government, and the public. To ensure this cooperation and coordination, all of these participants should commit to ... [a] set of common fundamental principles of CIP [1]".

Historically, control systems were implemented in stand alone environments where they were isolated from corporate networks and the Internet. Recent years have seen these systems becoming more and more interconnected; to corporate networks, vendors and support service provider networks and even to the Internet, using both wired and wireless communications technologies. These interconnections increase the security risk to process control systems claims Amin [11]. According to Amin, "... traditional external entities such as suppliers, consumers, regulators, and even competitors must now have access to segments of the network" thus rendering these key infrastructures highly vulnerable to either accidental or intentional failure. Because of the interconnections, single, isolated incidents "can cascade through and between networks with potentially disastrous consequences" [11]. Improved interconnectivity has also resulted in the centralising of process control operations, thereby making centralized control centres potential terrorist targets, not only for cyber attacks, but also for physical attacks [8].

Hardware, software and communications protocols implemented in the early process control systems comprised of proprietary technologies, and the knowledge of these technologies was limited to the 'engineers' and specialist operators. Specialist skills and knowledge was required to use these systems and use the information processed within these systems. The possibility of a malicious or inadvertent security breach on these systems was therefore highly improbable. However, there has been a recent trend where these proprietary technologies are being replaced with industry standard, common operating systems and applications that are vulnerable to common information security threats [9][10]. With the proliferation of information on the Internet, not only product documentation and technical specifications, but also how-to papers and easy-to-use exploitation tools on process control systems technologies have become freely available [9].

In addition to the technological vulnerabilities, there are also several operational vulnerabilities that affect process control systems. While proprietary technologies have been replaced with common industry standard technologies, some of the specialised features of process control systems and implementation idiosyncrasies of these systems have not been adequately addressed in the new implementations. While some implementations address the new process control environment as they did with the previously isolated environments, the others have adopted an approach more suitable for traditional IT infrastructures. This has resulted in various inherent design and operational vulnerabilities within these systems including poor security administration and insecure network connections. One of the most common issues related to this situation is the lack of policy specific to the process control environment. There is a distinct difference between IT and process control environments. Therefore the process control environment needs a separate and specific security administration and policy structure [12].

Perhaps one of the major challenges in developing protection plans for process control systems is posed by the diverse range of devices and software applications in use within the process control environment and their different implementations. The

author has audited process control systems in several energy and water utilities. While some of these utilities perform the same functions and carry out the same operations, there has been no two sites that used the same process control systems deployed in the same manner, even when they used the same service provide for systems integration and ongoing support. This poses a serious challenge in as each site will require a different approach to implementing and managing security of the process control systems.

There are many standards applicable to process control system security, and this adds to the complexity of securing and managing security on an ongoing basis within a process control environment. A 2005 research report by the US Department of Energy identified 18 standards or report findings applicable to securing process control systems in the Energy Industry [13]. The professional and regulatory bodies that provide these standards include:

- Institute of Electrical and Electronic Engineers (IEEE),
- US Federal Energy Regulatory Commission (FERC),
- National Electricity Reliable Council (NERC),
- International Electro Technical Commission (IEC),
- American Petroleum Institute (API),
- the Instrumentation, Systems and Automation Society (ISA),
- The National Institute for Standards in Technology (NIST), and
- International Organisation for Standardisation (ISO)

These standards attempt to address various aspects of process control system security. Some of these standards are industry specific while others are applicable to one or more industries and provide generic guidelines. This makes the selection of applicable standards to a particular process control environment an onerous task as organisations need to not only identify and implement applicable standards, but also manage the implementation of these standards on an ongoing basis.

4 Managing the Security of the Process Control Systems – The Need for a Standardised Approach

While there are a number of standards and methodologies already in place to address the security aspects of process control systems, there is not a single streamlined approach available to assist organisation to select and implement standards applicable to their environment. There are a number of security standards applicable to process control systems defined by the North American Electricity Reliability Council (NERC). For example, NERC has defined a series of standards to provide guidance in protecting the critical infrastructure within the Electricity Industry consisting of the following:

- CIP-002-1 – Critical Cyber Asset Identification;
- CIP-003-1 – Security Management Controls;
- CIP-004-1 – Personnel and Training;
- CIP-005-1 – Electronic Security Perimeter;
- CIP-006-1 – Physical Security;

- CIP-007-1 – Systems Security Management
- CIP-008-1 – Incident Reporting and Response Planning; and
- CIP-009-1 – Recovery Plans for Critical Cyber Assets;

In addition to these security standards there are industry specific standards that address operational aspects of the industry, which if not adhered to can result in the failure of the critical services. For example, NERC identifies the following industry specific standards in the electricity generation, transmission and distribution industry:

- Resource and Demand Balancing;
- Critical Infrastructure protection;
- Emergency Preparedness and Operations;
- Interchange Scheduling and Coordination;
- Interconnection Reliability Operations and Coordination;
- Modelling Data and Analysis;
- Protection and Coordination;
- Transmission Operation; and
- Transmission Planning

NERC have defined 132 Standards within these generic and specific categories that define the reliability requirements for planning and operating the North American bulk electric system. These standards can be found online at <https://standards.nerc.net>. While some of these standards address reliability of electricity operations, the underlying infrastructures rely to a great extent on process control systems for efficient control and monitoring of operations.

Most of these standards are normative, and rely on the organization to determine how they are interpreted and applied. For example NERC cyber security framework is defined in a series of standards labelled CIP-001 through CIP-009. As NERC states:

“NERC Standards CIP-002 through CIP-009 provides a cyber security framework for the identification and protection of Critical Cyber Assets to support reliable operation of the Bulk Electric System.

These standards recognize the differing roles of each entity in the operation of the Bulk Electric System, the criticality and vulnerability of the assets needed to manage Bulk Electric System reliability, and the risks to which they are exposed. Responsible Entities should interpret and apply Standards CIP-002 through CIP-009 using reasonable business judgment [14]”.

There are several key issues related to the approach taken by NERC with these standards.

Firstly, each standard will require a process to identify the applicability, select and implement controls to comply with the standard. This can become a burdensome task to the organization.

Secondly, a heavy reliance on individuals to use ‘reasonable business judgement’, as some definitions of the responsible entities identify individuals as the responsible entity. This leaves room for errors on judgement and a non-standard approach to the security of process control systems.

Thirdly, these standards do not have an ongoing performance measurement and improvement mechanisms to address the dynamics of the security lifecycle. With the

large number of standards involved, it also becomes almost impossible to implement an audit programme to ensure ongoing conformity of the standards.

To address these issues, it is necessary to identify the relationship between these standards and treat the application of these standards as parts of a larger system. It is also necessary to implement a management system that takes a holistic, streamlined and dynamic approach that has both auditable performance metrics and a process for continuous improvement built in. It is also imperative that the responsibility for the risk assessment and decision making is returned to the management of the organisation.

To ensure a comprehensive approach aligned with business direction and organisational risk management programs, the ISO/IEC 27001 standard is hereby proposed as the Information Security Management Systems framework for process control systems.

5 Information Security Management Systems (ISMS) Standards

The ISO/IEC standards on information security management consist of two complementary standards. These two standards are:

1. ISO/IEC 27001:2005 - Information Technology - Security Techniques - Information Security Management Systems – Requirements (ISO 27001)
2. ISO/IEC 17799:2005 - Information Technology - Code of Practice for Information Security Management (ISO 17799)

ISO/IEC 17799 provides the non-normative code of practice for information security management. It details the generally accepted practices in Information security management, and defines and describes the components of an Information Security Management System (ISMS). ISO 27001 contains the normative, auditable information security management standard. ISO 27001 defines the set of requirements for the implementation of an ISMS that needs to be met if an organisation requires formal certification and accreditation against the standard. It contains a methodology for the development of an ISMS and provides a collection of security controls that an organisation can select from based on the security risks and requirements of that organisation.

The approach used in the ISO 17799 and 27001 is based on risk management. ISO 17799 contains an introductory clause on risk assessment and treatment, followed on by 11 control clauses identifying key areas of risk and security practices that could be used to address those risk areas. These control clauses are as follows:

- Section 4 – Risk Assessment and treatment
- Section 5 – Security policy
- Section 6 – Organising information security
- Section 7 – Asset management
- Section 8 – Human resources security
- Section 9 – Physical and environmental security
- Section 10 – Communications and operations management
- Section 11 – Access control

- Section 12 – Information systems acquisition, development and maintenance
- Section 13 – Information security incident management
- Section 14 – Business continuity management
- Section 15 – Compliance

Compliance against sections 4-8 is mandatory for organisations seeking ISMS certification.

There are 11 control clauses, 39 control objectives and 134 individual controls addressing a comprehensive range of potential risks. The standards identify that individual organisations may have unique security requirements and provides the flexibility to select controls applicable to their organisation, or implement additional controls not included in the standard. However, there needs to be risk assessment based management justification for the omission of a control. Appendix A of the ISO 27001 standard contains direct one-to-one mapping of sections 4-15 of the ISO 17799 standard in a normative format.

The ISO 27001 Standard follows a process approach to manage to establish, implement, operate, monitor, review, maintain and improve the organisation's ISMS. All ISMS Processes are structured under a "Plan-Do-Check-Act" (PDCA) model which provides the basis for understanding the information security requirements, addressing the organisational information security risks, monitoring and reviewing the performance and the effectiveness of the ISMS and the continual improvement of the processes and the ISMS.

6 Applicability of the ISO 27001 ISMS to the Process Control System Environment

To demonstrate the applicability and benefits of implementing an ISMS framework to manage, monitor and improve information security of process control systems, the NERC CIP Framework will be used as an example.

A simple one-to-one mapping of the CIP standards with the ISO 17799 standard is demonstrated in table 1.

A brief comparative analysis indicates that the ISO 17799 ISMS standard contains controls to address all areas addressed by the NERC CIP series of standards in a single manageable framework.

For some organisations, it may be a requirement that compliance requirements with the NERC CIP standards are met. The ISO 17799 ISMS framework can be used to simplify such compliance requirements. As ISO 27001 offers the flexibility of incorporating controls outside of Annex A, it would be possible to incorporate the CIP series of Standards as part of the ISMS for a process control system, thereby addressing the requirements of both the NERC and ISO 27001 standards.

One of the key benefits of implementing using the ISO 17799 ISMS framework to implementing would be the additional safeguards that it adds to the NERC CIP standards. The ISMS Framework is periodically audited for compliance; this means that the CIP standards that comprise the ISMS controls would maintain their currency and relevance.

Table 1. Mapping of NERC CIP Standards to ISO 17799 Controls

NERC CIP Standards		ISO 17799 Controls	
Risk Assessment	CIP-003-1 – Security Management Controls;	Section 5 – Security policy	Section 4 – Risk Assessment and Treatment
		Section 6 – Organising information security	
	CIP-002-1 – Critical Cyber Asset Identification;	Section 7 – Asset management	
	CIP-004-1 – Personnel and Training;	Section 8 – Human resources security	
	CIP-006-1 – Physical Security;	Section 9 – Physical and environmental security	
	CIP-005-1 – Electronic Security Perimeter;	Section 10 – Communications and operations management	
	CIP-007-1 – Systems Security Management		
		Section 11 – Access control	
		Section 12 – Information systems acquisition, development and maintenance	
	CIP-008-1 – Incident Reporting and Response Planning; and	Section 13 – Information security incident management	
	CIP-009-1 – Recovery Plans for Critical Cyber Assets;	Section 14 – Business continuity management	
		Section 15 – Compliance	

Another key benefit is that the ownership of the ISMS is retained by the management, providing the management with complete visibility of the risks associated with the process control systems and the critical infrastructure. The management forum for information security required by the ISMS would ensure that all decision making related to the process control systems receives the organisational input, and would no longer be reliant on ‘reasonable business judgement’ of an individual.

Section 15 of the ISO 17799 standard addresses compliance with legal and other regulatory requirements. Controls within this section can be used to address the other operational compliance requirements such as NERC Emergency Operations Standard EOP-003 – Load Shedding Plans.

The ISMS also offers complete lifecycle management to the security process. Within the process driven PDCA lifecycle, the security of the process control systems will continuously be monitored with input from the organisation for continual improvement of security.

7 Conclusion

By implementing an ISO/IEC 27001 information security management system, the organization adopts a comprehensive and systematic approach to the security of the process control systems. The coordinated risk assessment and treatment plan that

extends the focus of the risk assessments from a single system or process to an organisational risk assessment, taking into consideration all factors that can affect the systems or the process including inter and intra organisational interdependencies. The result would be the assurance that not only the individual systems, but also the integrity of complex larger system of which individual process control systems participate in.

References

1. Attorney General's Department (Australia): Trusted Information Sharing Network for Critical Infrastructure Protection, Attorney Generals Department, 2006
2. Attorney General's Department (Australia): Critical Infrastructure Protection National Strategy. Canberra, Attorney General's Department, 2004
3. Dacey RF: Critical Infrastructure Protection: Challenges in Securing Control Systems, United States General Accounting Office, 2003
4. Rockliff M: Process Control System Security, Plexal Group, 2005
5. Verton D: Blaster Worm Linked to Severity of Blackout, ComputerWorld, ComputerWorld, 2003
6. Poulson K: Slammer Worm Crashed Phio Nuke Plant Network, SecurityFocus, 2003
7. Boyer SA: SCADA Supervisory Control and Data Acquisition, 3rd Edition. Research Triangle Park, NC, ISA - The Instrumentation, Systems and Automation Society, 2004
8. Miller A: Trends in Process Control Systems Security. IEEE Security and Privacy 2005; 3: 57-60
9. US Computer Emergency Readiness Team: Control Systems Cyber Security Awareness, US-CERT, 2005
10. Byres E, Lowe J: The Myths and Facts behind Cyber Security Risks for Industrial Control Systems, British Columbia Institute of Technology, 2004
11. Amin M: Infrastructure Security: Reliability and Dependency of Critical Systems. IEEE Security and Privacy 2005; 3: 15-17
12. Kilman D, Stamp J: Framework for SCADA Security Policy. Albuquerque, Sandia National Laboratories, 2005
13. Carlson R, Dagle JE, Shamsuddin SA, Evans RP: A summary of Control System Security Standards Activities in the Energy Sector, Department of Energy, 2005, pp 48
14. North American Electricity Reliability Council: Reliability Standards for the Bulk Electric Systems of North America, 2006

A Composite Key Management Scheme for Mobile Ad Hoc Networks

Yingfang Fu¹, Jingsha He², and Guorui Li¹

¹ College of Computer Science and Technology
Beijing University of Technology
Beijing 100022, China
{fmsik, liguorui}@emails.bjut.edu.cn
² School of Software Engineering
Beijing University of Technology
Beijing 100022, China
jhe@bjut.edu.cn

Abstract. An Ad Hoc network is a multi-path autonomous system comprised of a group of mobile nodes with wireless transceivers. In this paper, we first review the present research in key management in Ad Hoc networks. Then, by considering the characteristics of Ad Hoc networks in which the free movement of nodes can lead to a variety of topological changes, especially network separation and convergence, we propose a new key management scheme based on a combination of techniques, such as hierarchical topology structure, virtual CA (certification authority), offline CA and certificate chain. We show that the proposed scheme would improve key management in security, expandability, validity, fault tolerance and usability.

Keywords: Ad Hoc networks, key management, identity-based cryptosystem, threshold cryptography, hierarchical cluster algorithm.

1 Introduction

A mobile Ad Hoc network is a wireless network with the characteristics of self-organization so that it can quickly form a new network without the support of wired network infrastructure. Mobile Ad Hoc networks not only play an important role in military applications, but also have a wide variety of commercial applications such as disaster recovery. On the other hand, compared with the wired and normal wireless networks, mobile Ad Hoc networks are more prone to security threats and attacks, e.g., passive interception, data interpolation and replaying, identity forgery, and denial of service. That is because mobile Ad Hoc networks usually operate in a wide open space and their topologies change frequently, which makes such networks lack of centralized mechanisms for protection and clear lines for defense. Therefore, security is essential for applications in mobile Ad Hoc networks. Currently, most of the research in mobile Ad Hoc networks can be found in two areas: routing protocols and key management [1] However, many routing protocols for Ad Hoc networks are designed with the assumption that all communication between mobile nodes is secure, which is not true in

general because there are a lot of potential threats in such a wireless environment, e.g., modification of routing protocols and forgery of IP addresses. Consequently, security has become the central issue for mobile Ad Hoc networks and key management plays a central role in network security for ensuring confidentiality, integrity and non-repudiation of information. In this paper, we propose a new key management scheme using such techniques as hierarchical topology structure, virtual CA, offline CA, certificate chain and proactive measurement. We show that the proposed scheme would improve key management in security, expandability, validity, fault tolerance and usability.

The rest of this paper is organized as follows. In the next section, we review some related work in key management schemes for Ad Hoc networks and the hierarchical cluster algorithm. In Section 3, we describe the factors that should be considered in the development of key management schemes in mobile Ad Hoc networks. In Section 4, we present a new key management scheme along with a thorough analysis. Finally, in Section 5, we conclude this paper with a discussion on our future work

2 Related Work

We review some previous work in key management in mobile Ad Hoc networks in this section. We also discuss the hierarchical cluster algorithm as it is a fundamental technique to be used in our scheme.

2.1 Key Management Schemes

In general, we don't assume that there is a trustworthy node in a mobile Ad Hoc network. Therefore, we don't presumably have a single node to act as the Certificates Authority (CA) in key management. In this section, we describe some key management schemes and point out their limitations.

Partial Distributed Key Management Scheme. Zhou proposed a threshold key management scheme based on public key encryption [2] The method uses a trusted offline organization and the (k, n) -threshold scheme to protect the private key. The offline organization would issue a certificate to a mobile node when it joins the network and generate a public key and a private key for the system, in which the public key is distributed to all the mobile nodes in the network. The secret of the private key is, however, shared by n serving nodes, which are randomly selected in the network by the offline organization. The n serving nodes would then manage the certificates according to the (k, n) -threshold scheme [2] Therefore, the network can withstand the situation in which at most $k-1$ serving nodes are compromised without compromising the system security with respect to key management. Each serving node has the public key of every other mobile node in the network and can thus communicate with all the nodes securely. But this scheme has the following shortcomings:

- (1) Since the method randomly selects n nodes as the serving nodes, it could make the system less stable [3, 4]
- (2) Since each serving node stores the public keys of all the other mobile nodes, it would need a large amount of memory space.

- (3) Some distributed CA schemes don't provide the means for revoking certificates.
- (4) The method doesn't have corresponding means to deal with network separation caused by the mobility and loss of nodes. Therefore, the method can hardly support network expandability.
- (5) When the network gets very large, it would be difficult, if not impossible, for the serving nodes to get the public keys of all the other nodes. Consequently, this scheme is not suitable for large networks.

Self-organizing Key Management Scheme. The self-organizing key management scheme proposed by Hubaux et al [5, 6] uses a certificate chain graph to implement key management. There is no centralized CA and every node creates its own private key. The distribution of the public key certificate is purely the business of the node itself. For two nodes, say, nodes A and node B, if A trusts B, A could issue a certificate to B that contains B's public key encrypted with A's private key. In addition, there is no centralized certificate directory and every node has its own certificate directory that contains a list of certificates including those that the node has sent to other node, those that the node has received from other nodes, and those that have been created using the Maximum Degree Algorithm [5] The trust relationship between nodes can be described using a certificate chain graph. In the graph, a vertex represents the public key for a node while an edge represents a certificate issuance for the public key. Assuming that K_A and K_B are the public keys of node A and node B, respectively, if A has issued a certificate to B, there would be a directed edge from A to B in the certificate chain graph. If two nodes want to certify each other, they would merge the two certificate directories and look for the certificate chain in the certificate chain graph that connects A and B. The method doesn't have any special requirement on the mobile nodes and every node is completely independent. Therefore, it is more suitable for Ad Hoc networks with the characteristics of self-organization. Nonetheless, the method has the following shortcomings:

- (1) The certificate chain that connects two nodes may not be found.
- (2) Without a trusted organization in the system to provide guarantee, the method would make it less trustable for a long certificate chain.
- (3) Since the method depends on the trust relationship between nodes, vicious node may destroy the network security through forging large numbers of certificates.
- (4) The method doesn't provide the means for revoking certificates.

Identity-based Key Management Scheme. The identity-based key management scheme proposed by Khalili et al [8] uses node identification and system public key to implement key management, in which the system public key is generated and broadcast in the network by n special nodes, namely the PKG nodes. The PKG nodes also issue certificates using the (k, n) -threshold secret sharing scheme [5]. Node identification relies on unique IDs such as node names, postal addresses or Email addresses. A node doesn't have a specific public key and would obtain a private key from PKG nodes In order to obtain a private key, a node needs to contact at least k PKG nodes, provides its identity and related information, gets partial private key from each such node, and

compose its own private key. The node identity consists of a random string of characters, which would make it unpredictable, and the PKG nodes issue an exclusive private key for every node. The method could reduce the complexity of computation as well as communication cost because each node would not have to create its own public key and broadcast it in the network. But the method has following two shortcomings:

- (1) A malicious node may pretend to be a PKG node and forge a public key to a new node, while possessing a corresponding private key.
- (2) The method doesn't provide any means to deal with key update.

2.2 Hierarchical Cluster Algorithm

To deal with node mobility and improve the stability of cluster frame, we could assign a weight value to each node based on its mobility. The more frequently a node moves, the lower its weight value will be. A cluster head is the node with the largest weight value. In the hierarchical cluster algorithm [8], node mobility is expressed in terms of relative mobility by comparing the strengths of two signals that the node receives from its neighbors. In the algorithm, the relative mobility index of node y relative to node x using formulas (1) below:

$$M_y(x) = 10 \log \frac{Rxpr_{x \rightarrow y}^{new}}{Rxpr_{x \rightarrow y}^{old}} \quad (1)$$

in which $Rxpr_{x \rightarrow y}^{new}$ represents the power that node y receives from node x and $Rxpr_{x \rightarrow y}^{old}$ represents the power that node y received from node x last time. If $M_y(x) < 0$, it means that the distance between two nodes becomes longer, otherwise it becomes shorter. Through computing the mean value of the absolute values of the relative mobility between node y and all the other neighboring nodes x_i ($i \in m$), The local mobility value of node y can be obtained using formulas (2) below:

$$M_y = \frac{\sum_{i=1}^m |M_y(x_i)|}{m} \quad (2)$$

The less the value M_y is, the lower the relative mobility of the node in relative to all the other neighboring nodes.

3 Criteria for Key Management Schemes in Mobile Ad Hoc Networks

A mobile Ad Hoc network is a communication network with the capability of self-organization. In such a network, mobile nodes can move within the network as well as join in and drop out of the network freely, which can cause frequent changes to

network topology. Therefore, we should consider the following factors in the development of key management schemes:

- (1) What is the specific application for which a mobile Ad Hoc network is used? This is because different applications may require different levels of security.
- (2) When mobile nodes join in or drop out of the network, or are destroyed, how can we detect these events in real-time and perform key management appropriately?
- (3) Network-level decision making usually requires the cooperation of more than one mobile node. Then, what is the cooperation strategy that can help to protect the network from malicious attacks aimed at the cooperation process?
- (4) How can we extend the security mechanism when the network changes continuously?
- (5) How can we guarantee the fault tolerance, feasibility and efficiency of the security mechanisms?

In the next section, we will propose a composite key management scheme for mobile Ad Hoc networks based on the above five criteria.

4 The Composite Key Management Scheme

4.1 The Network Model

The whole network consists of one or more sub-networks. We call each such sub-network a cluster. Each cluster will have a head node, $n-1$ PKG nodes and many other ordinary cluster nodes. The cluster heads in turn form the next layer of the network, and so on. Nodes can roam from one cluster to another. Nodes in a lower layer network, e.g., in a cluster, communicate within a relatively shorter range and those in a higher layer network communicate within a relatively longer range. Let's assume

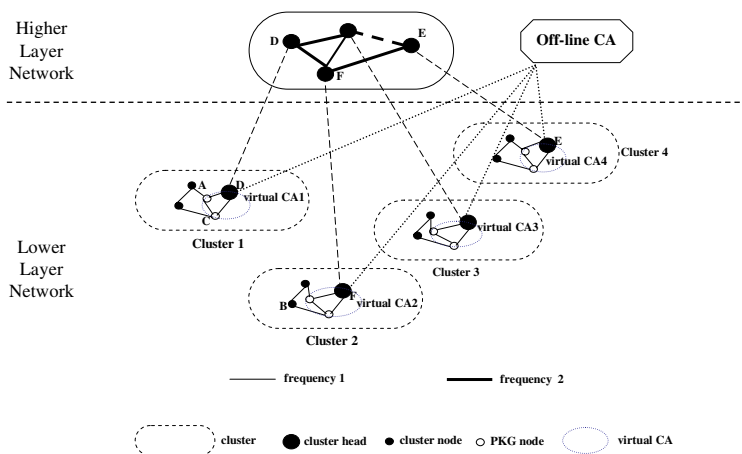


Fig. 1. The Network Model

without loss of generality that cluster nodes communicate with each other in frequency 1, the cluster heads communicate with each other in frequency 2, and so on There is an off-line CA in the network that appoints the cluster heads. The CA also performs functions such as keeping track of the status of the network, detecting topology change and collecting reports from cluster heads For each cluster, the cluster head selects $n-1$ cluster nodes as the PKG nodes that are generally high performance nodes in terms of security, data storage and computation power. These PKG nodes would then form a virtual CA along with the cluster head. The cluster head is responsible to detect topology change within the cluster as nodes join and leave the network or are destroyed. When necessary, the cluster head would make a connection to the off-line CA although most of the time they are not connected Fig. 1 illustrates the network model in our key management scheme.

4.2 Key Creation

Key creation is illustrated in Fig. 2, which involves nine steps:

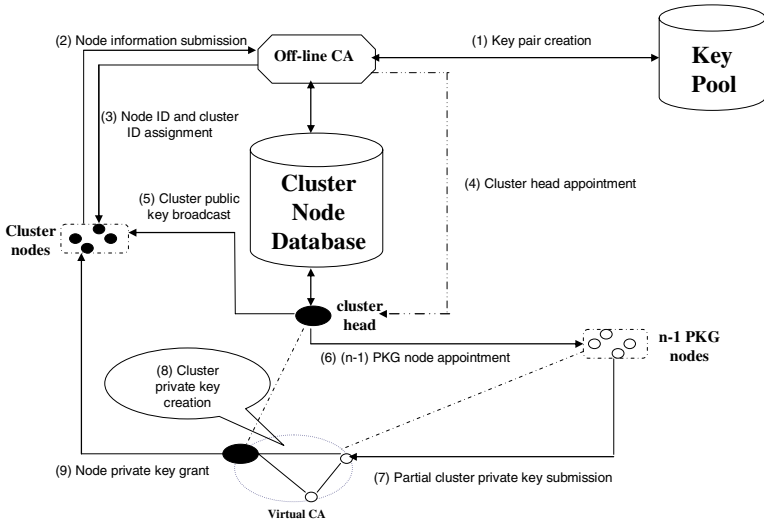


Fig. 2. Key Creation

- (1) The off-line CA creates a pair of private and public keys for each cluster and generates a certificate.
- (2) Before joining the network, every cluster node must register with the off-line CA.
- (3) The off-line CA needs to identify a new cluster node and then assigns it a cluster ID and node ID.
- (4) The off-line CA selects one cluster node from among all the cluster nodes according to the hierarchical cluster algorithm and appoints it as the cluster head. The cluster head is also granted the right to access the cluster nodes information database.

- (5) The cluster head broadcasts the public key for the cluster to all the other cluster nodes. The cluster head selects $n-1$ cluster nodes as the PKG nodes. PKG nodes are generally high performance nodes in terms of security, data storage and computation power. The cluster head then forms a virtual CA together with the $n-1$ PKG nodes.
- (6) The cluster head and the $n-1$ PKG nodes manage the keys using the (k, n) -threshold cryptographic method. That is, the private key for the cluster is partitioned and held by the cluster head and the $n-1$ PKG nodes. To reconstruct the private key, any k out of the $n-1$ PKG nodes would have to submit partial private keys to the cluster head.
- (7) The cluster head could compute a new private key for the cluster based on its own private key, the old private key for the cluster and all the partial cluster private keys submitted to it by the k PKG nodes.
- (8) The cluster head would create the private key for a cluster node based on the cluster private key and grants it to the cluster node.

4.3 Key Update and Revocation

Cluster nodes can roam from one place to another in the network, resulting in the need that an existing key has to be replaced by a new one. There are two reasons for it. One is that when a cluster node joins the network, it should not be allowed to access information transmitted in the past. The other is that when a cluster node drops out of the network, it should not be allowed to access information to be transmitted in the future any further

To detect the abnormal behavior of a node in real time, which occurs when a new node joins the network, an existing node drops out of the network or is destroyed, we use a cluster-based mobile intrusion detection subsystem that runs on every cluster node to monitor and collect information about the status of every other node in the cluster as well as that of the system through intrusion detection mechanisms [9, 10]. There are typically five modules in such a subsystem: the data collection module (DCM), the intrusion detection engine module through analysis of patterns (IDEMP), the intrusion detection engine module through analysis of abnormality (IDEMA), the local association and aggregation module (LAAM) and the global association and aggregation module (GAAM). In such a subsystem, DCM is responsible for collecting local trace information and journal information. IDEMP and IDEMA are responsible for detecting abnormal behavior of nodes through analyzing the data collected by DCM. More specifically, IDEMP detects abnormal behavior of nodes by comparing the behavior of nodes with patterns of attack and IDEMA detects abnormal behavior of nodes by comparing the behavior of nodes with the descriptive normal behaviors of nodes. LAAM associates and aggregates the results produced by IDEMP and IDEMA. Lastly, the function of GAAM is determined by the type of the cluster node in which the GAAM executes; in a normal cluster node that is not a cluster head, GAAM would send the results to the GAAM of its cluster head, while in a cluster head, GAAM would associate and aggregate its own results with those provided by other cluster nodes and would then send the aggregated results to the offline CA.

When the off-line CA is notified of node's joining in, dropping out or being destroyed in a cluster, it would connect to the network and update the keys depending

on the following different scenarios. In the first scenario, the change is caused by an ordinary cluster node. The off-line CA would then update the key pair for the cluster with the cluster head. The cluster head would then announce the newly joined node to all the cluster nodes and update the private key for the cluster as well as that for each cluster node together with the PKG nodes. In the second scenario, the change is caused by a PKG node. Then, in addition to updating the private key for a cluster, the cluster head must select new PKG nodes from the remaining cluster nodes and would then update the private keys accordingly. In the third scenario, the change is caused by the cluster head. Then, the off-line CA not only updates the key pair for the cluster, but also selects a new cluster head from the remaining cluster nodes. The newly selected cluster head would then select $n-1$ PKG nodes from the remaining cluster nodes and update the keys accordingly.

To reconstruct the private key, the cluster head would have to work with at least k PKG nodes. To prevent attacks, the off-line CA, together with the virtual CA, would update the private key for each cluster as well as for each cluster node periodically. Therefore, each and every private key has a timestamp associated with it, which limits the validity of the private key.

4.4 Communication Between Cluster Members

The communication mechanism must ensure the security of any communication between cluster nodes, between cluster heads, and between the cluster head and cluster nodes.

Communication Between Cluster Heads. Communication between cluster heads could take place under two circumstances with one being that cluster heads fully trust each other and the other being that they don't. If two cluster heads trust each other, one cluster head can directly send to the other cluster head a public key certificate created with its private key. Every cluster head maintains its own certificate directory that contains those certificates that it has granted to other cluster nodes, those that it has received from other cluster nodes, and those that have been created using the Maximum Degree Algorithm [5].

Communication Between Cluster Nodes. When two cluster nodes in the same cluster communicate with each other, one node would encrypt messages using the public key of the cluster and the ID of the other cluster node. The messages can only be decrypted by the other cluster node because it owns the corresponding private key. This also applies to the situation, in which the cluster head communicates with a cluster node that is in the same cluster. When a cluster node wants to communicate with another cluster node that is not in the same cluster, there are usually five steps involved:

- (1) The sending cluster node would ask its cluster head for the public key of the cluster to which the receiving cluster node belongs.
- (2) When the cluster head of the sending cluster node receives the request message encrypted with the public key of the cluster and the ID of the sending cluster node, it decrypts the message with its private key, and would then obtain the ID of the receiving cluster node.

- (3) The cluster head would manage to get the public key for the cluster to which the receiving cluster node belongs by asking the off-line CA or certificate chain based on the ID of the receiving cluster node, and would forward it to the sending cluster node.
- (4) When the sending cluster receives the message that its cluster head returns to it, it decrypts the message with its own private key, and then obtains the public key of the cluster to which the receiving cluster node belongs.
- (5) The sending cluster node can now communicate with the receiving cluster node with the public key of the cluster to which the receiving node belongs and the ID of the receiving cluster node securely.

5 Conclusion

In this paper, we presented a composite key management scheme for mobile Ad Hoc networks. By analyzing and identifying the advantages and limitations of existing key management schemes in such networks, we can see that the scheme that we proposed has the following advantages:

- (1) It uses a hierarchical topology structure, which can be easily extended to deal with mobile Ad Hoc networks of any sizes.
- (2) Every cluster has its own public key, which is known only to all the cluster nodes in the same cluster and all the other cluster heads. A new cluster private key is computed based on the old cluster private key and the partial private keys created by the k PKG nodes. The cluster public key, the cluster private key and the private key of cluster nodes all have the timestamp in them. These characteristics help to improve the validity and security in key management.
- (3) The off-line CA connects to the mobile Ad Hoc network only under certain circumstances, which improves not only the trustworthiness of key management, but also its security.
- (4) By using identity-based cryptographic means and the threshold scheme, cluster nodes don't need to create and broadcast their own public keys in the network, which helps to improve performance by taking less network bandwidth and storage space.
- (5) When cluster heads communicate with each other, keys can be managed with the certificate chain technique and the off-line CA, which improves key management in the areas of fault tolerance, serviceability and availability. For a small network in which cluster nodes trust each other, the certificate chain technique would make the cluster heads communicate more conveniently. For a large network in which cluster nodes don't necessarily trust each other, the off-line CA not only improves the trustworthiness among the cluster heads, but also helps to overcome the limitation of the certificate chain.
- (6) The cluster heads and cluster nodes keep track of and collect status of the clusters with proactive measurement techniques, which could improve key management in dealing with continuous topological changes.

The scheme that we presented in this paper only represents our initial effort for the development of a sophisticated key management scheme for mobile ad hoc networks. As the future work towards achieving our eventual goal, we need to develop and

improve the intrusion detection subsystem used in our scheme based on analysis of intrusion patterns as well as on behavior abnormality. In addition, although we have argued qualitatively that our composite key management scheme is more advantageous over several previous schemes in terms of message exchange overhead and performance, we need to do more quantitative analysis through simulation or measurement to further justify our claim. Consequently, we will need to improve the proactive measurement technique used in our scheme and perform a thorough performance evaluation on our composite key management scheme through simulation.

References

1. Cai, H.J.: The Security Hole and Strategy for Mobile Ad Hoc Network Microcomputer Development, Vol. 14 (2004)
2. Zhou, L. and Haas, Z.J.: Securing Ad Hoc Network. *IEEE Networks*, Vol. 13, No. 6 (1999)
3. Yi, S. and Kravets, R.: Key Management for Heterogeneous Ad Hoc Wireless Networks. In Proc. 10th IEEE International Conference on Network Protocols (2002)
4. Yi, S. and Kravets, R.: MOCA: Mobile Certificate Authority for Wireless Ad Hoc Network. In Proc. 2nd Annual PKI Research Workshop (2003)
5. Hubaux, J.P., Buttyan, L. and Capkun, S.: Self-Organized Public-Key Management for Mobile Ad Hoc Networks. *Transactions on Mobile Computing* (2003)
6. Hubaux, J.P., Buttyan, L. and Capkun, S.: The Quest for Security in Mobile Ad Hoc Networks, *CACM* (2001)
7. Khalili, A., Katz, J. and Arbaugh, W.A.: Toward Secure Key Distribution in Truly Ad-Hoc Networks. In Proc. Symp. on Applications and the Internet Workshop (2003) 342-346
8. Zheng, S.R.: The Technique of Ad Hoc Network, Posts & Telecom press (2004)
9. Jiang, H.: Intrusion Detection Technologies in Mobile Ad Hoc Network. *ZhongXing Telecom Technology*, Vol. 10 (2002)
10. Whit, MGB, Fisch, EA, and Pooch, UW.: Cooperative Security Managers: A Peer-Based Intrusion Detection System *IEEE Network*, Vol. 10, No. 1 (2001) 20-30

Adaptive Algorithms to Enhance Routing and Security for Wireless PAN Mesh Networks

Cao Trong Hieu¹, Tran Thanh Dai¹, Choong Seon Hong², and Jae-Jo Lee^{3,*}

¹ Department of Computer Engineering, Kyung Hee University
Giheung, Yongin, Gyeonggi, 449-701 Korea
{hieuct, daitt}@networking.khu.ac.kr
<http://networking.khu.ac.kr>

² Department of Computer Engineering, Kyung Hee University
Giheung, Yongin, Gyeonggi, 449-701 Korea
cshong@khu.ac.kr

³ Korea Electrotechnology Research Institute
665 Nesonong, Euiwang, Gyeonggi, Korea, 437-808 Korea
jjlee@keri.re.kr

Abstract. Wireless PAN Mesh Network (WMN) is currently going to be standardized and enhanced to take full advantages of the flexible and heterogeneous networks. Although the standard (802.15.5) is under-construction, WMNs are expected to become popular as they have the ability to connect all kinds of current networks. So far, there is no applied architecture which is efficient enough to completely solve routing and security problems in WMN. To assist IEEE P805.15 in routing and security aspects, in this paper, we propose an adaptive algorithm for detecting bogus nodes when they attempt to intrude into the network by attacking routing protocol. In addition, a procedure to find the most optimal path between two nodes is presented along with adaptive pre-conditions for WMNs. We also show that our algorithm is robust according to the mobility of the nodes and it is easy to implement in currently proposed architectures. It can work with many kinds of wireless networks as well as can reduce computational costs.

Keywords: Wireless PAN Mesh Networks, Security, Intrusion Detection, Clustering, Optimal Path, Routing, Attack on Routing Protocol.

1 Introduction

Wireless Mesh Network (WMN) could be considered as a successor of the basic wireless networks such as Wireless LAN, Wireless Mobile Ad-hoc Networks (MANETs) and Wireless Sensor Networks (WSNs), which inherits full advantages of the previous ones and could be applied in many fields in daily life as well as in military operations that require dynamic topology. However, WMNs also require to deal with

* This work was supported by MIC and ITRC Project. Dr. C.S. Hong is the Corresponding Author.

inherent weaknesses of wireless networks by consequence of dynamic topologies and node mobility. Moreover, the lack of concentration points where traffic can be analyzed for intrusions leads to utilize self-configuring multi-party infrastructure protocols that are susceptible to malicious manipulation and subject to noise and intermittent connectivity due to the inherent essence of wireless communication channels [7], [8], [9].

In WMNs, the topology is a mixture of star (with access point as coordinator) and grid (ad-hoc based) and therefore the routing protocol must be flexible for adaptive change. In our scenario, the coordinators can be mobile nodes (mobile access points or ad-hoc nodes) or station (fixed access points like WLAN access points). In this paper, we not only focus on solving routing problem of mobile nodes but also propose a mechanism to automatically work with fixed access points.

The rest of the paper is organized as follows, Section 2 briefly discusses some related works as well as addresses assigning problem which adapts to WMNs, in Section 3, we propose a procedure to find the most optimal path between two nodes when they want to communicate with each other. This procedure can be applied in any kind of wireless dynamic network. Moreover, we propose an algorithm for detecting bogus nodes when they attempt to intrude into network by attack routing protocol. Section 4 presents our simulation results. Finally, section 5 exposes some perspectives for further work.

2 Related Work

2.1 Current Work

Clustering technique is proposed in many papers for routing and formation of a dynamic topology. In security, it is also used to detect intruders. Based on the clustering technique, which was first proposed by Zhang and Lee [17], D. Sterne in [5] has given an architecture in which the author solved most of the drawbacks of a dynamic topology when implementing an Intrusion Detection System (IDS). However, as almost related papers, the author just only explained some clues to detect bogus node in routing protocol that are used to conclude that they can in principle determine whether a node is an attacker. Moreover, the authors did not specify the attributes of a node to decide whether it has enough qualifications or not to find intermediate nodes in routing protocol.

Although the proposals above are applied in MANETs, they can also be applied in the Wireless Mesh Networks that are with almost same dynamic topology by making some slight modifications to routing protocol. As we know, the attack in routing protocol is very hard to prevent, especially in the wireless environment where the traffic can be easily eavesdropped, therefore we improve current contributions by taking advantages of previous proposals and using them as building blocks for our proposals. In this paper, AODV (Ad hoc On-demand Distance Vector) [18], [19], an adaptive protocol, is also utilized in our work.

Currently, the IEEE P802.15 Working Group for WPANs has been making a standard for WMNs with many achievements [1], [2]. In those proposals, they also used mesh tree, another form of cluster, to solve almost problems in routing. However,

they did not focus on security for routing, as well as giving a solution to finding optimal path between two nodes. Our proposal is a contribution for completing the 802.15.5 standard in aspect of routing and security.

2.2 Address Assigning Problem

Cluster-tree technique can be applied in any kind of network that has dynamic topology. In the cluster-tree, a node can have a maximum number of C_M children and a node can be at most L_M levels (i.e., mobile devices) away from the root of the tree (C_M and L_M are two predetermined network-wide constants). A node with a short address s is in charge of assigning short addresses to its children as in the following algorithm [16]: assign short address $s+1$ to the first child, $s+1+C_{hold}(L_N)$ to the second child, and $s + 1 + (n - 1) \cdot C_{hold}(L_N)$ to the n th child, up to the (C_M) th child. And $C_{hold}(L_N)$ is calculated as follows:

$$C_{hold}(L_N) = \left\lfloor \frac{T - \sum_{k=0}^{L_N} (C_M)^k}{(C_M)^{L_N+1}} \right\rfloor$$

$$T = \sum_{k=0}^{L_M} (C_M)^k$$

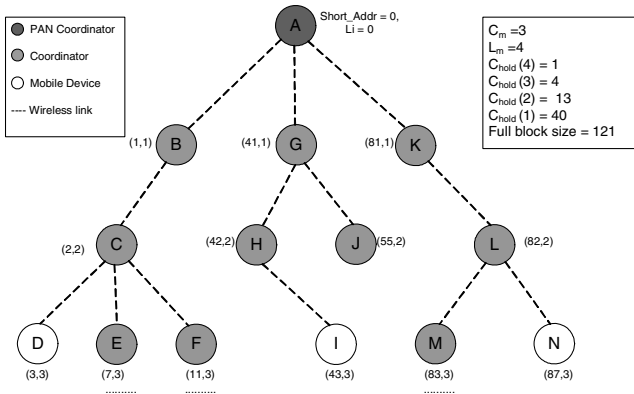


Fig. 1. Assigning Address

In which:

T : Address Block Size

$C_{hold}(L_N)$: Number of address each node of specific level can hold

L_N : Level of Node

C_M : Maximum number of children of a specific level node

L_M : Maximum level in a cluster-tree

After executing some mechanisms to establish and maintain the dynamic topology as well as to choose the coordinator, this coordinator begins to accept association requests from other nodes. Any node already existent in the network can determine whether to allow other nodes to join to the network, that is, whether to act as a coordinator, depending on the availability of its resources such as memory and energy. In a cluster-tree, a node is able to calculate the next hop by looking at the destination address in the packet. This precludes the need of route discovery, and thus helps reduce the initial latency, control overhead, memory usage and energy consumption.

In the Figure 1, an example of assigning address is given with $L_M=4$, $C_M=3$. It means the coordinators in level3 only have maximum 3 children (mobile terminals), and the rest of the remaining nodes can be assigned their own address as well as their holding address block size which rely on the level of nodes.

3 Proposed Algorithms

The hybrid topology of WMNs makes their routing problem more difficult than homogeneous networks. In this case, we need a routing protocol that can work in two network structures: Mobile nodes with fixed Access Point (as Coordinator) and entirely mobile nodes (Ad-hoc topology)

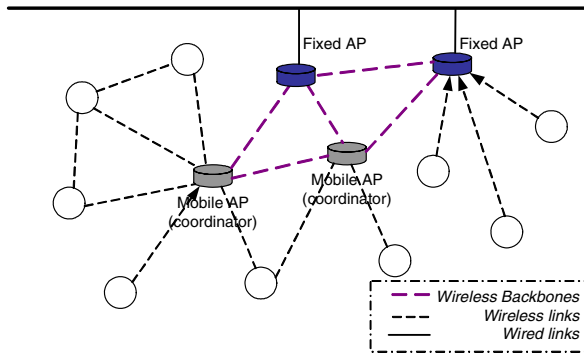


Fig. 2. Mesh Topology

In the previously proposed cooperative intrusion detection architecture using clustering technique [5], the authors presented their solutions to detect attack on routing protocol, but they did not give any algorithm to prove that their technique can detect and exclude bogus nodes. Furthermore, they did not also give any procedure with specific criteria to find the shortest (the most optimal) path.

In this paper, we do two jobs: the first one is proposing a procedure to find a shortest (the most optimal) path using three most important criteria, signal strength, bandwidth and energy remaining. The other is proposing mechanism to identify and exclude bogus nodes.

To implement our proposal, some pre-conditions are established.

We use the clustering technique to establish and maintain the dynamic hierarchy according to node mobility. We also use AODV (Ad hoc On-demand Distance Vector) as routing protocol in our proposal. In addition, for hybrid structure in WMNs, we propose some small changes in routing protocol:

1. When two nodes in the ad-hoc part of WMNs want to communicate with each other, they use Multicast Routing Protocol [1] with our adaptively proposed procedure presented below.
2. When one node in ad-hoc part wants to communicate with others in fixed Access Point (AP) part, it finds the AP where the destination is currently connect with and after that can transfer data through this AP.
3. When two nodes in the fixed AP part of WMNs want to communicate with each other, their current APs will act as intermediate nodes and use our proposed procedure to find shortest path.

3.1 Optimal Path Finding Algorithm

There are many criteria to decide whether a node has ability and capacity to become an intermediate node in a route. In such a dynamic topology like WMNs, it is very difficult to find a completely good routing protocol which can automatically reform and maintain connection. The most three important criteria we use in our procedure are *Signal strength*, *Bandwidth*, and *Energy Remaining* because they guarantee for a stable and high-speed connection. When a node wants to communicate with another one, the following steps are processed:

Step 1: Initial RREQ = 1, BroadcastID = 1, the source node floods RREQS packets with destination address to its neighbors and chooses the node with the most powerful signal strength.

Step 2: Estimate the available Bandwidth and Energy Remaining of this node, if its free bandwidth $\geq 50\%$ & If necessary time \geq Data-Size/Bandwidth

Choose this node as a next hop

Else, repeat *Step 1* to choose another node, remember information of current node to compare with new found node to find the most optimistic node. B.CastID++;

End if

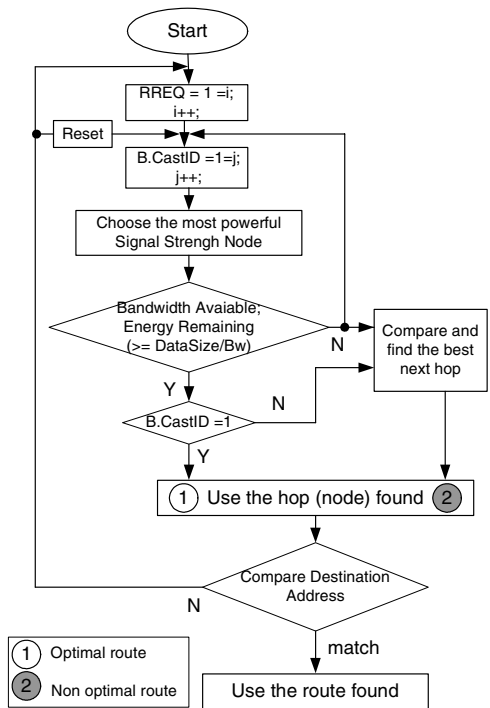


Fig. 3. The Most Optimize Path Procedure

Step 3: If Node's B.CastID = 1
 Use the found node
 Else, use this node to compare with previous found node and choose the best one.
Step 4: Compare Destination Address
 If found Node's Address is the same as Destination Address, go to *Step 5*
 Else { repeat *Step 1*; Reset B.CastID = 1; }
Step 5: Choose the route
 Finish

In Step 2, the available bandwidth is assigned $\geq 50\%$. This value can be adjusted to suit requirement in a specific network. If a node with the highest signal strength and enough bandwidth is found in each hop, along with enough energy remaining, it means the optimal route is found. If not, we can also find the best route at the final part of Step 3. Available bandwidth and energy remaining can be easily estimated in nodes themselves with current softwares.

In Step 3, the sender has known the size of the packet that it intends to transmit to receiver. In addition, with currently available bandwidth (can be evaluated by each node itself in the route), the necessary time can be calculated and compared with the remaining energy time of each node in the route. Based on the requirements of network, we can add other criteria such as proximity, resistance to compromise, accessibility, processing power, storage capacity, etc. to the procedure.

If a node in the route satisfies all conditions in the procedure, that node is an optimal one. If a route has all optimal intermediate nodes, it is called optimal route.

3.2 Identify and Exclude Bogus Nodes

This algorithm is used to detect and exclude intruders at any time they attempt to break routing mechanism.

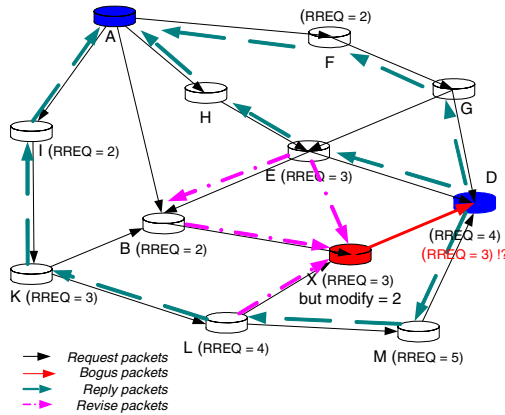


Fig. 4. Identify Bogus Node

Initial

$RREQ_S = 1, RREP_D = 0, Found\ Route\ Count = 0;$

flood $RREQ_S$ in the network topology;

for each time $RREQ_S$ reaches node i

do $RREQ_i = RREQ_i + 1;$

$N_{hop\ i} = RREQ_i;$

compare destination address;

if destination found

do Co-revise Procedure

{

for each route found from S to D

Found Route Count++;

send $RREP_D$ back through other routes different from route of the first reach $RREQ$ packet;

$RREP_{Dj} = N_{hop\ j} - j;$

compare ($RREQ_j, RREP_j$) index;

if $RREP_j$ index determined by neighbor nodes $\neq RREQ_j$ index;

trigger an alarm;

}

exclude j out of connection;

Finish

The number of routes found is counted by Found Route Count and is stored in routing table to help routing protocol distinguish different routes, decide which routes are chosen and keep the information for optimal path finding as well as back-up purpose.

To avoid bogus nodes from modifying RREP packets before sending them back to the same route, the destination node will send RREP packets through other routes. By this way, Co-revise Procedure can completely identify intruders. The number of backward routes is limited to avoid redundancy and reduce the number of nodes involved in routing procedure. At least two neighbors of bogus node X will ensure that X is intruder, after comparing RREQ index that X modified with its real RREQ by the following revising mechanism:

Assume X is attacker and it is trying to access to the route between A and D . Following the procedure, $RREQ_A = 1$, A floods its request to find optimal route to D . In the figure 4, $RREQ_{B, H, F, I} = 2$ because they are neighbors of A , and $RREQ_{G, E, K} = 3$ and so on until the RREQ reaches D . If X is a legal node and it is in network topology, $RREQ_X$ must be 3, but it modified this index, suppose $RREQ_X = 2$, and sends to D . In principle, D will “think” the route include X is optimal, and choose this route.

But now D can use proposed algorithm above, send back RREP to the other routes, like D-E-H-A and D-G-F-A. After that, B and E can themselves calculate the real RREQ index of X, and find it have to logically equal 3. Also, if $RREQ_X = 2$, it means X have to be a neighbor of A like B,E,F, but A can itself determine C is not a neighbor because A can not directly communicate with X. In brief, the algorithm can definitely detect X is intruder, trigger an alarm and exclude X out of network.

In the Figure 4, we can see path A-I-K-L-M-D is also exist, but destination node D will not use it because the number of intermediate nodes is large. The routing protocol in [1] has already solved this problem. Therefore, D will not send back RREP_D through this path and thereby limiting the number of nodes have to involve in routing protocol. This mechanism also help our proposals save energy, reduce time consumption and memory usage.

4 Simulation Result

Our proposed algorithm is simulated to further evaluate the theoretical results. We use OMNeT++ Ver.3.2 with Mobility Framework 1.0a5 Module. We present each node as a matrix in which attributes (*Signal Strength, Band Width, Energy Remaining, Address, etc.*) are assigned as factors. We use routing table in [1] with additional fields *Found Route Count* and *Sequence Number*. We also set up a mobility environment to evaluate the performance in detection rate and calculation time influenced by different movement speed of nodes. The nodes have radio range of 300m and move on the rectangular surface according to the boundless mobility model. We study the detection rate, cardinality and time consuming according to mobility and network cardinality.

In the figure 5, the detection rate of bogus nodes is 99.05% for a set of 25 nodes at speed 5m/s and. The detection rate is a little bit decrease according to the increase of

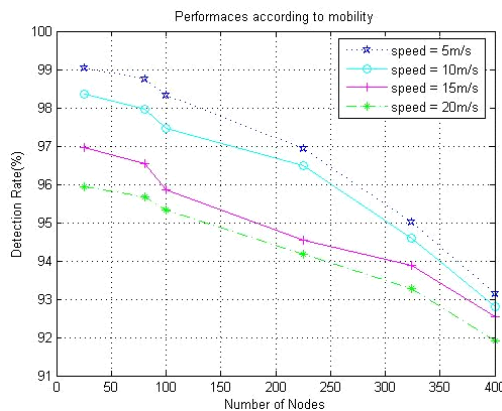


Fig. 5. Performance according to mobility

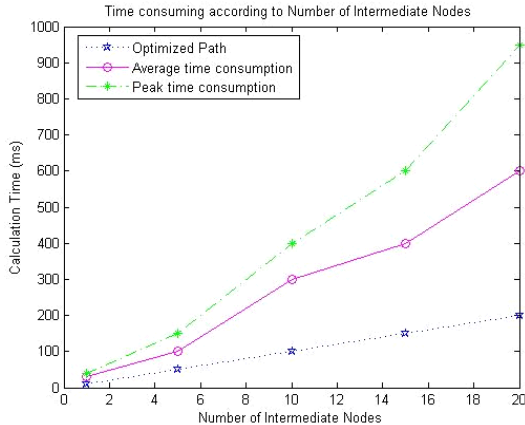


Fig. 6. Time consuming

number of nodes as well as speeds. At speed 20m/s and 400 nodes, the detection rate is reduced to 91.92%.

In figure 6, the time consuming for finding intermediate nodes in case of optimal path is also the least. It is directly proportional to the numbers of intermediate nodes. We evaluate the peak of calculation time and average time in case of non-optimal path.

5 Conclusion and Future Work

Our proposed approach in this paper bases on dynamic topology maintained by clustering technique, uses AODV as the routing protocol, inherits the achievements of previous researchers and improve shortcomings in their proposals. By making adaptive changes, our algorithm can be applied to any kind of Wireless Network such as WMNs, MANETs and WSNs. To apply our algorithm, we just insert additional fields Found Route Count and Sequence Number into routing table. The simplicity of our algorithm is that it does not require a considerable amount of computational resource, even there are a large number of nodes in a selected route. Each time the algorithm find the next hop, the process returns to the initial point at Step 1 and does the same jobs until the destination is found. Consequently, the number of times that needed to process is direct proposition with the number of intermediate nodes in route, and the complexity in each Step is trivial.

Moreover, our proposals can work with currently used protocols and completely solved routing problem for nodes in different wireless networks.

In future works, we will continue implementing our proposal in Testbed cooperating with current Intrusion Detection Systems (IDSs) for Wireless Networks. Furthermore, we are working on an algorithm for automatic reforming topology based on clustering technique which will run in company with our proposals.

References

1. IEEE 802.15-15-05-0247-00-0005, "Mesh PAN Alliance (MPA)", IEEE 802.15.5 Working Group for Wireless Personal Area Networks.
2. IEEE 802.15.5-05-0260-00-0005, "IEEE 802.15.5 WPAN Mesh Network", IEEE 802.15.5 Working Group for Wireless Personal Area Networks.
3. Guan, Y. Ghorbani, A. Belacel, "A clustering method for intrusion detection". Proceedings of Canadian Conference on Electrical and Computer Engineering, 1083-1086, Canada, 2003.
4. Stefano Basagni, "Distributed Clustering in Ad Hoc Networks", Proceedings of the 1999 Intl. Symp. On Parallel Architectures, Algorithms and Networks (I-SPAN '99), Freemantle, Australia, 1999.
5. S. D.Sterne, "A General Cooperative Intrusion Detection Architecture for MANETs", Proceeding of the Third IEEE International Workshop on Information Assurance (IWIA'05), 0-7695-2317-X05 IEEE, 2005.
6. Yi-an Huang, Wenke Lee, "A Cooperative Intrusion Detection System for Ad Hoc Networks", 2003 ACM Workshop on Security of Ad Hoc and Sensor Networks (SASN '03) George W. Johnson Center at George Mason University, USA, October 31, 2003.
7. Vesa Karpijoki, "Security in Ad Hoc Networks" <http://citeseer.nj.nec.com/karpijoki01security.html>
8. P. Brutch and C. Ko, "Challenges in Intrusion Detection for Ad Hoc Network", IEEE Workshop on Security and Assurance in Ad hoc Networks, Orlando, FL, January 28, 2003.
9. Konrad Wrona, "Distributed Security: Ad Hoc Networks & Beyond", PAMPAS Workshop, London, Sept. 16/17 2002.
10. F. Theoleyre, F. Valois, "A Virtual Structure for Hybrid Networks", IEEE Wireless Communications and Networking Conference (WCNC 2004), Atlanta, USA, March 2004.
11. P. Krishna, N. H. Vaidya, M. Chatterjee, and D. K. Pradhan, "A cluster-based approach for routing in dynamic networks", ACM SIGCOMM Computer Communication Review, 1997.
12. M.-Y. Huang, R. J. Jasper, and T. M. Wicks, "A large scale distributed intrusion detection framework based on attack strategy analysis", Computer Networks, pp. 2465–2475, 1999.
13. Jake Ryan, Meng-Jang Lin, Risto Milikkuilainen, "Intrusion Detection with Neural Networks", Advances in Neural Information Processing Systems 10 (Proceedings of NIPS'97, Denver, CO), MIT Press, 1998.
14. K. Ilgun, R. A. Kemmerer, and P. Porras, "State transition analysis: A rule-based intrusion detection approach", IEEE Trans on Software Engineering, pp. 181–199, 1995.
15. P. A. Porras and P. G. Neumann, Emerald, "Event monitoring enabling responses to anomalous live disturbances", in Proc of 20th NIST-NCSC Nat'l Info Systems Security Conf, pp. 353–365, 1997.
16. Jianliang Zheng, Myung J. Lee, Michael Anshel, "Towards Secure Low Rate Wireless Personal Area Networks", IEEE Transactions on mobile computing.
17. Y. Zhang and W. Lee, "Intrusion Detection in Wireless Ad Hoc Networks", Proceedings of The Sixth International Conference on Mobile Computing and Networking (MobiCom 2000), Boston, MA, August 2000.
18. Charles E. Perkins, Elizabeth M. Belding-Royer, and Samir Das. "Ad Hoc On Demand Distance Vector (AODV) Routing", IETF RFC 3561.
19. C. E. Perkins and E. M. Royer, "The ad hoc on-demand distance-vector protocol", In C. E. Perkins, editor, Ad Hoc Networking. Addison-Wesley, 2000.

Secure and Seamless Handoff Scheme for a Wireless LAN System

Jaesung Park¹, Beomjoon Kim², and Iksoon Hwang³

¹ Department of Internet Information Engineering, The University of Suwon,
Gyeonggi-Do, 445-743, Korea
jaesungpark@suwon.ac.kr

² Department of Electronic Engineering, Keimyung University, Daegu, 704-701 Korea
bkim@kmu.ac.kr

³ Core S/W 1 Team R&D) LG-Nortel, Gyeonggi-Do, 431-749, Korea
iksoonhwang@lg-nortel.com

Abstract. IEEE 802.11i standard specifies full authentication and preauthentication for secure handoff in 802.11 wireless LAN (WLAN). However, the full authentication is too slow to provide seamless services for handoff users, and preauthentication may fail in highly populated WLANs where it is highly probable that the cache entry of a preauthenticated user is evicted by other users before handoff. In this paper, we propose a seamless and secure handoff scheme by reducing authentication and key management delay in the handoff process. When a user handoffs, security context established between the user and the previous access point (AP) is forwarded from the previous AP to the current AP, and the session key is reused only for the handoff session. The freshness of session key is maintained by regenerating session keys after handoff session is terminated. The proposed scheme can achieve considerable reduction in handoff delay with providing the same security level as 802.1X authentication by letting an AP authenticate a handoff user before making an robust security network association (RSNA) with it.

1 Introduction

The wireless local area networks (WLAN) based on IEEE 802.11 infrastructure-mode have been deployed successfully as an economical means to provide users ubiquitous broadband access to Internet. Unlike cellular networks where users can handover while having on-going calls, WLAN systems have provided only portability where users can move only within the radio coverage of an access point (AP) to which they are connected. That is, users cannot move while using the network because current WLAN systems do not easily support seamless handoff. However, as users experience with wireless network increases, they demand continuous communication while on the move. Therefore, fast handoff becomes one of the important research issues in the evolution steps of WLANs.

In a WLAN, handoff initiated by a mobile node (MN) goes through the following 4 logical steps: probing, reassociation, authentication and key creation.

First, MN seeks to find potential next APs in the probing phase. After making a handoff decision, MN reassociate with an AP to which it decides to handoff. Then MN is reauthenticated by a network and new session keys are generated between MN and network for the handoff session.

Security is as important as fast handoff for successful WLAN deployment because data is transferred via wide open wireless radio. However, authentication process involves a few message exchanges between MN and an authentication server (AS) in a network which is generally located far away from APs. It also takes a few interactions between AP and MN to create new session keys for handoff session. The long delay for security on WLAN becomes the major obstacle that makes fast handoff difficult.

To solve the delay problem in authentication, pre-authentication is included in the 802.11i specification [1]. Basically, preauthentication try to avoid reauthentication by authenticating each MN to a set of potential next APs before it handoffs to one of them. However, 802.11i does not specify how to select a set of candidate APs. Several researchers try to answer this question. Frequent handoff region (FHR) is proposed in [2] to denote the set of potential next APs with the long term movement history of MN. Neighbor graph is proposed to determine the potential set of APs [3]. They note the number of candidate APs is small fraction of the total APs. However, these proactive methods must be carefully engineered to avoid reauthentication. For example, the security context of a MN_i in a candidate AP could be updated by the other MNs before the MN_i handoffs to the AP. It is quite probable if the density of the MNs in the coverage of an AP is high and they move frequently, which is the case of WLAN system deployed in hot spots. If the security context is not found when a MN handoffs, a full authentication process takes place to fail to support seamless service. Also, proactive scheme is not scalable because it imposes heavy management loads on a single AS and each APs with a large signaling messages between them.

In this paper, we propose a reactive solution which supports the same security level as IEEE 802.11i specification in terms of authentication and freshness of session key while reducing handoff delay significantly. We focus on reducing key creation delay after handoff as well as the authentication delay. We augment the 802.11i specification to implement the proposed method for backward compatibility. Specifically, we add two fields in the reassociation request message in IEEE 802.11 MAC management frame and one field in the capability information to make an AP authenticate the MN requesting the secure reassociation without involving an AS. When an MN handoffs from AP_i to AP_j , the security context of the MN installed at AP_i is fetched to AP_j . Using the context information and reassociation request frame, AP_j can authenticate the MN requesting reassociation. Also, we reuse the temporary key created before handoff only for the termination of the handoff session. However, the freshness of session key is maintained by regenerating session keys after handoff session is terminated. Unlike proactive schemes, our method operates consistently regardless of network environments such as density of mobile nodes and their movement pattern without incurring heavy management overhead of an AS.

2 Fast and Secure Handoff Problems in WLAN Systems

In this section, we explain typical WLAN network architecture and best current practice for secure WLAN based on IEEE 802.11i specification. With the discussion, we derive the problems caused by security mechanisms in providing fast handoff. We also review related works for fast and secure handoff to discuss their advantages and disadvantages.

2.1 WLAN Security Based on IEEE 802.11i

In terms of IEEE 802.11i specification, secure WLAN is defined as robust security network (RSN) where all mobile nodes and APs make robust security network association (RSNA) between them. RSNA is made when MN and AS authenticates each other and MN and AP generates a temporary secure key for data encryption over wireless link. To build a RSN, IEEE 802.11i specifies authentication enhancement based on IEEE 802.1X over entity authentication such as open system authentication and shared key authentication. It also specifies key management and establishment, encryption enhancement over wired equivalent privacy (WEP). In the 802.11i, it is assumed that the AS and the AP to which a mobile station associates is trusted. Moreover, it is implicated that AS and APs have trust relationship. In a typical WLAN system which is owned and operated by a single carrier, network management tools are provided to detect unauthorized APs, therefore trust relationship between APs can be assumed.

When a MN handoffs in a RSN, it must establish RSNA with a new AP again. That is, a MN must be authenticated again by an AS and temporary security key be created. For mutual authentication, extensible authentication protocol (EAP) is used between a MN and an AS. EAP allows a MN to select specific authentication method such as EAP-TLS, EAP-MD5, EAP-AKA, however, EAP-TLS [4] is often used. EAP-TLS messages are exchanged between a MN and an AP over wireless link encapsulated by EAP over LAN (EAPoL) protocol. IEEE 802.11i does not mandate protocols between APs and an AS. However, remote authentication dial-in user service (RADIUS) becomes a de facto standard. Recently, EAP over DIAMETER is being developed. After mutual authentication, a session key for data encryption over wireless link is created through IEEE 802.11i protocol called four way handshake.

EAP-TLS provides challenge-response type strong authentication and encryption. For the EAP-TLS authentication, MN and AS must have certificate from common certification authority (CA). Figure 1 shows the complete message flows during authentication and four way handshake. Authentication process starts by sending the identity information of a MN to AS. Then, a MN authenticates AS via AS certificate. After successful authentication, MN randomly select a pre-master secrete and send the premaster secrete encrypted with the public key of the AS (Client-Key-Exchange message) to the AS with its certificate. The AS can authenticate the MN with its certificate. With the premaster secrete both

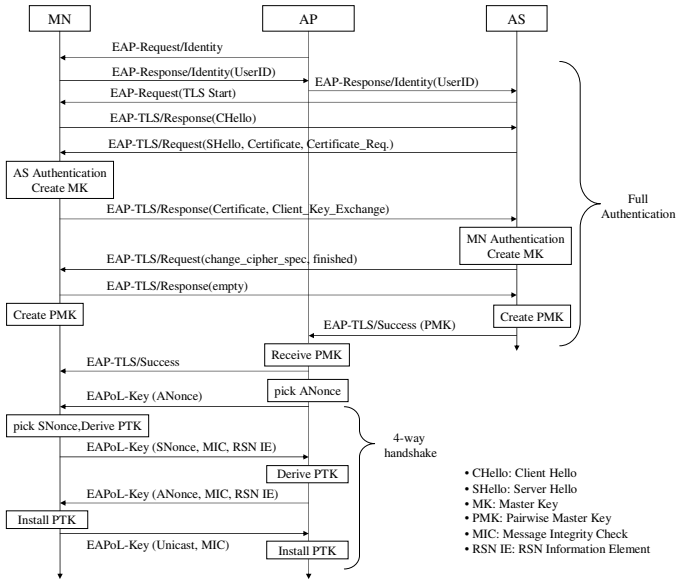


Fig. 1. Full authentication and 4-way handshake procedure

the MN and the AS creates a master key (MK). The MK is used to derive a pairwise master key with a pseudo random function (PRF) as follows.

$$PMK = PRF(MK, CHello || SHello). \tag{1}$$

The AS sends PMK to an AP which the MN requests to make a RSNA. Therefore, after successful mutual authentication, MN and AS share the MK, and MN, AS, and AP have the common PMK. PMK is used to generate a pairwise transient key (PTK) for data encryption between MN and AP. Four way handshake using EAPoL-Key messages takes place to confirm the liveness of the MN and AP, and to guarantee the freshness of the PTK. MN and AP exchanges its randomly selected Nonce (ANonce from AP, SNonce from MN) through the first two EAPoL-Key messages. PTK is created using the PMK and medium access control (MAC) addresses of the MN and AP as well as ANonce and SNonce by using the following equation.

$$PTK = PRF(PMK, MN_{MAC} || AP_{MAC} || ANonce || SNonce). \tag{2}$$

The third EAPoL-Key message is used to synchronize the PTK between MN and AP and the fourth message signifies the completion of the four way handshake and the installation of the key.

2.2 Preauthentication Schemes

From the above discussion, it is apparent that full authentication and four way handshakes are major obstacles for fast handoff because they require a number

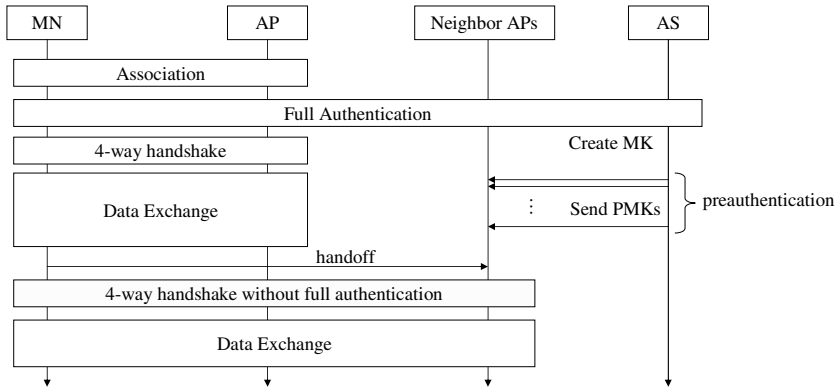


Fig. 2. Preauthentication procedure

of message exchanges among MN, AP, and AS which takes an order of seconds. To solve the problem, preauthentication is also included in the IEEE 802.11i specification. Basically, preauthentication try to reduce handoff delay by authenticating each MN to a set of potential next APs before it actually handoffs to one of them. Figure 2 illustrates the message flows when preauthentication is used.

However, 802.11i does not specify how to select a set of candidate APs. Several researchers try to answer this question. Pack [2] proposed a frequent handoff region (FHR) to denote the set of potential next APs. The FHR of a MN is calculated from the long term movement history of the MN. A centralized AS records and analyzes the frequency each MN moves from one AP to another. That is, AS maintains $n \times n$ matrix of each MN, where n is the number of AP in the WLAN system and the element of the array N_{ij} is the inverse of handoff ratio of the MN from the AP_i to AP_j . If the MN associates with an AP_i , it also authenticates with other APs in FHR. Neighbor graph (NG) is proposed to determine the potential set of APs [3]. They note the number of candidate APs is small fraction of the total APs. Neighbor graph can be constructed in a distributed manner at each AP or it can be installed in an AS when WLAN is deployed. The latter is often used for its fast convergence time. Once neighbor graph is established, an AS distributes the security context and key materials of an MN to the set of APs in the neighbor graph. If an MN moves to one of the candidate APs in the neighbor graph, the authentication process is avoided.

However, these proactive schemes have the following drawbacks. First, the performance of the scheme depends not only on the prediction mechanisms but also on the cell environment. For example in a neighbor graph scheme, the security context of a MN_i in a candidate AP could be updated by the other MNs before the MN_i handoffs to the AP. It is quite probable if the density of the MNs in the coverage of an AP is high and they move frequently which is the case of WLAN system deployed in hot spots. If the security context is not found when a MN handoffs, a full authentication process takes place which fails to

support a seamless service. Second, proactive schemes are not scalable in terms of state information maintained in a centralized AS and the signaling overhead between an AS and APs.

3 Proposed Secure and Seamless Handoff Scheme

In this section, we detail our seamless and secure handoff method. We extend IEEE 802.11i specification to implement the proposed method for backward compatibility. The fundamental idea is to authenticate handoff MN by a new AP with the previous security context from the old AP without involvement of an AS. Also, the PTK generated between MN and old AP is reused to eliminate the 4-way handshake delay, only for the duration of the handoff session. New PMK is delivered from AS to the AP while handoff session continues. Whether PTK expires or handoff session terminates, new PTK is created between new AP and MN to guarantee the freshness of session key.

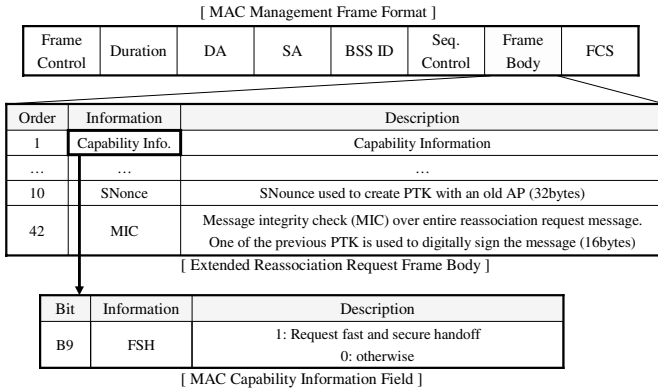


Fig. 3. Extension of reassociation request MAC management frame

3.1 Extended MAC Management Frame Body Components

To implement the proposed scheme, we extend the reassociation request MAC management frame body and capability information field as figure 3. To indicate the ability of secure and seamless handoff, MN sets the fast and secure handoff (FSH) bit in the capability information index within the reassociation request message. FSH bit is also included in the beacon message, probe response message, and association request message to indicate the ability of secure and seamless handoff of AP and MN. In the reassociation request frame body, handoff MN includes the SNonce it used to generate PTK with an AP to which it associates before handoff. MN also includes message integrity check (MIC) calculated over the reassociation request frame using the PKT. The AP to which the MN handoffs can check the integrity of the reassociation request frame using the MIC.

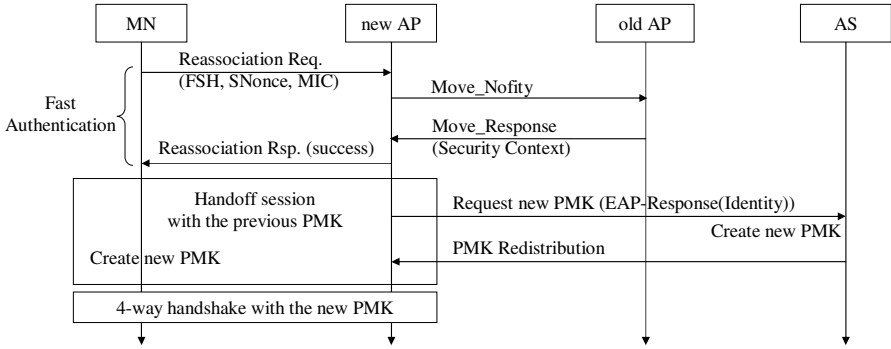


Fig. 4. Message flows in the proposed fast and secure handoff method

3.2 Fast Authentication and Key Management

Figure 4 illustrates the message flows of our proposed method. When a MN handoffs from old AP to new AP, MN sends a reassociation request message to new AP. New AP fetches the security context (e.g. PMK, SNonce, ANonce, MAC address of the old AP, Cipher Suite, etc.) of the MN from old AP. The inter-access point protocol (IAPP) can be used to exchange the security context of a MN between APs because it is developed to exchange information between APs from different vendors [5]. The only difference in our approach is that security association is assumed to be made between APs at the network deployment. However, the adaptation of IAPP is straight forward.

Because the MIC in the reassociation request message is encrypted by PTK used between MN and old AP, only the new AP can have the same PTK and cipher suite to correctly decrypt the MIC from the security context. If the message integrity check passes, new AP authenticates the MN by comparing the SNonce in reassociation request message and the SNonce in the security context from old AP.

After successful authentication, MN and new AP keep using the previous PTK only for the handoff session to reduce 4-way handshake delay. In the 802.11i, handoff is considered the same as the initial access to the network. However, we argue that handoff is the continuation of the on-going session because MN will use the same session key if the MN does not handoff. Therefore, MN is allowed to use the previous PTK if MN is authenticated by new AP. However, the reuse of the PTK must not sacrifice the freshness of session key and the liveness of the communicating entities. Session key must be refreshed after each session. For this purpose, new PMK is distributed from AS to new AP while handoff session continues. In the 802.11i trust assumption, new PMK must be different from the previous PMK. We devise the derivation of the new PMK which binds MK and MAC address of the old AP and new AP as follows.

$$nPMK = PRF(MK, oldPMK || MN_{MAC} || oldAP_{MAC} || newAP_{MAC}). \quad (3)$$

MK is shared only between MN and AS after MN goes through full authentication at its initial network access. Old PMK is generated by MK and random number (CHello, SHello) generated by MN and AS, and old PMK is known only to MN, AS, and old AP. The freshness of session key and the liveness of each communicating party is guaranteed because the PTK is created again through 4-way handshake with new PMK when the handoff session terminates or the PTK ages off.

4 Performance Evaluation

In this section we analyze and compare the RSNA delay among full authentication, preauthentication and the proposed scheme. We define the RSNA delay as the sum of authentication delay and key management delay.

The delays between MN and AP, AP and AS, and AP and AP are denoted by t_a , t_d , t_{ap} , respectively. From figure 1, RSNA delay of the full authentication becomes $13t_a + 8t_d$.

In case of preauthentication, authentication is avoided if the security context is stored at an AP to which a MN handoff. Otherwise, full authentication takes place. The cache entry of the preauthenticated MN can be evicted by the other MNs while the MN resides in the current AP. We assume there are ρ MNs in cell area of each AP and the size of cache in each AP is N_c . From the fluid flow model [7], the aggregate rate of MNs crossing the cell boundary is given by

$$C = \frac{\rho v L}{\pi}, \quad (4)$$

where v is the average velocity of a MN and L is the size of location area. If we denote the cell residence time of a MN as t_{cr} and the cumulative distribution function of the cell residence time as $F(t)$, then the probability of cache miss (p_m) when MN handoffs to one of candidate AP becomes

$$p_m = Pr\left(\frac{\rho v L}{\pi} t_{cr} > N_c\right) = 1 - F\left(\frac{\pi N_c}{\rho v L}\right). \quad (5)$$

Therefore, RSNA delay of preauthentication is given by

$$P_d = 4t_a + p_m(13t_a + 8t_d). \quad (6)$$

From figure 4, RSNA delay of the proposed scheme depends on the t_{ap} because it avoids 4-way handshake. Then we can represent the RSNA delay as

$$R_d = 2t_{ap}. \quad (7)$$

4.1 Numerical Results

From the research on mobility model, cell residence time of a MN can be modeled using generalized gamma function [6]. That is, the probability density function of t_{cr} is modeled as

$$f(t_{cr}, a, b) = \frac{1}{b^a \Gamma(a)} t^{a-1} e^{-t/b}, \quad (8)$$

where a is a shape parameter, b is a scale parameter, and $\Gamma()$ is the Gamma function. The distribution becomes more concentrated, as a scale parameter becomes smaller. t_a is determined by 802.11 medium access control (MAC) protocol among contending MNs and the wireless link bandwidth. Therefore, there may be large variation in t_a if controlled management channel is not used. On the contrary, major contributor to t_d and t_{ap} are transmission delay. In a wired network, the transmission delay is stabilized and mainly depends on the hop count. In WLAN, adjacent APs are connected through a layer 2 switch or an access router, so they are one or two hops away from each other. Since AS is located at the core of a network, t_d is much larger than t_{ap} .

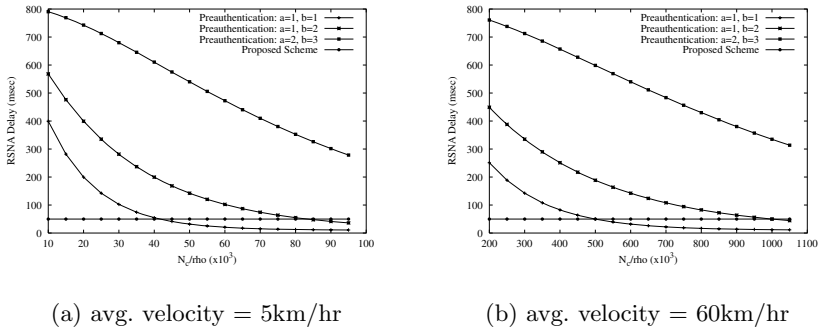


Fig. 5. RSNA Delay Comparison

Considering the latency budget for RSNA delay with 802.11b, which provides 11Mbps over wireless link, we assume $t_a=2.5$ msec, $t_d=97.2$ msec, $t_{ap}=23.7$ msec for numerical comparison between preauthentication and the proposed scheme. Figure 5 illustrates the RSNA delay for each scheme with different distribution of cell residence time of a MN when cell radius is 100m. We vary the average velocity of the other MNs from 5km/hr (figure 5-(a)) to 60km/hr (figure 5-(b)). The x-axis represents the ratio of the cache size of AP to the density of MNs within the radio coverage of an AP. As was noticed, preauthentication depends heavily on the ratio, the cell residence time of handoff MN, and the velocity of the other MNs. As the other MNs moves faster, APs need bigger cache to prevent the preauthenticated MN from being overwritten. Especially, when the variation in cell residence time becomes larger (for example, from $a=1, b=1$ to $a=2, b=3$), bigger cache is needed to cover the large deviation, which is not economical solution for deployment of many APs. On the contrary, the proposed scheme is affected only by the delay between APs and is not relevant to the N_c and the movement of the other MNs.

In terms of the management overhead, the proactive schemes need at least $O(n)$ computation and storage space per AP and AS for each MN, where n is the number of candidate APs per AP. Whereas, the proposed method only requires $O(1)$ computation and space per AP and AS, which makes it more scalable.

5 Conclusions

In this paper, we propose a reactive secure and seamless handoff method for WLAN system. The authentication delay is reduced by making a posterial AP authenticate MN requesting RSNA using the security context made with MN and previous AP. 4-way handshake is suspended until handoff session expires or PTK expires. We showed the proposed scheme is as secure as EAP-TLS authentication while reducing handoff delay. Compared to proactive method which depends on the other MN's mobility and the cell residence time of the handoff MN, our reactive method can bound handoff delay with a proper round trip time between APs without imposing heavy management loads both on APs and AS.

References

1. IEEE Std. 802.11i: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment 6: Medium Access Control(MAC) Security Enhancement, July (2004).
2. S. Pack and Y. Choi: Fast Inter-AP Handoff Using Predictive-Authentication Scheme in a Public Wireless LAN, *IEEE Networks*, Aug. (2002).
3. A. Mishra *et al.*: Proactive Key Distribution Using Neighbor Graphs, *IEEE Wireless Communications*, Feb. (2004).
4. B. Aboba, and D. Simon: PPP EAP TLS Authentication Protocol. RFC 2716, Oct. (1999).
5. IEEE Std. 802.11f: IEEE Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distributed Systems Supporting IEEE 802.11 Operation, July (2003).
6. M. M. Zonoozi and P. Dassanayake: User Mobility Modeling and Characterization of Mobility Patterns, *IEEE JSAC* vol. 15, no. 7, Sept. (1997).
7. H. Xie *et al.*: Dynamic Location Area Management and Performance Analysis, *Proc. VTC '93*, May (1993).

A CAPTCHA in the Text Domain

Pablo Ximenes¹, André dos Santos¹, Marcial Fernandez², and Joaquim Celestino Jr.²

¹ UPRM – Mayagüez Campus (RUM) – ZIP: 00680 – Mayagüez – PR – USA

² Av. Paranjana, 1700 – CEP: 60.740-000 – Fortaleza – CE – Brazil

pablo@ximenes.info, andre.moura@ece.uprm.edu,

{marcial, celestino}@larces.uece.br

Abstract. Research on CAPTCHA has led CAPTCHA design into adopting almost exclusively graphical implementations that deal mostly with character recognition. This has reached an exhaustion point, where new approaches are vital to the survival of the technique. This paper discusses the early stages of a research that intends to solve the open problem of a CAPTCHA in the text domain offering, this way, innovative research possibilities to the CAPTCHA paradigm. It is essentially an investigation on a CAPTCHA that draws its security from the cognitive and computational aspects behind phonetic punning riddles found on Knock-Knock Jokes. By the specification of a computational model, the implementation of a prototype and its experimentation with human individuals, it is shown that the proposal is indeed feasible and that studies in non conventional areas for Information Security are the key for developing the proposed goal.

Keywords: Security, CAPTCHA, Text Domain, Natural Language, Computational Humor.

1 Introduction

An important class of attacks within the internet requires the use of automated procedures to acquire access, privileges, or to exhaust resources of a system. For example, attacks that have as a goal to harvest free e-mail accounts require the use of automated tools to be able to register a large number of fabricated users.

One of the main strategies in defeating general automation based attacks is the CAPTCHA [1] paradigm. This technique implements a type of filter to computational systems that allows human access while denying service to computational robots, thus preventing automated attacks from taking place. In order to achieve that, CAPTCHAs explore several problems within the field of Artificial Intelligence (AI). One could argue that, because of their AI related characteristics, CAPTCHA systems would range within several interesting types of implementations, each one exploring different areas of AI. This does not happen, though. In fact, CAPTCHA implementations are mainly of graphical nature and deal mostly with pattern recognition, most specially with character recognition. This lack of innovative tracks has led CAPTCHA research into an exhaustion point, where new approaches are not only necessary but vital to the survival of the technique.

One interesting and innovative approach, regarded as an important open problem by the very CAPTCHA research community, is the problem of a CAPTCHA in the text domain. A CAPTCHA of this sort would draw its inner workings from AI problems that could be found within text constructs and would require only plain text to be assembled. Besides granting new possibilities to the CAPTCHA paradigm, a CAPTCHA in the text domain would be specially suitable for devices with low accessibility capabilities and for visually impaired users, since it would not require any advanced graphical screens and no multimedia features.

This way, in order to alleviate the forementioned problems, we propose a novel type of CAPTCHA that functions within the text domain. This paper describes the early stages of our research on the development of such CAPTCHA. In essence, the proposed CAPTCHA draws its security from the cognitive and computational aspects related to phonetic punning riddles found on Knock-Knock Jokes and similar structures. By the specification of a computational model, the implementation of a prototype and its experimentation with human individuals, it is shown that the proposal is indeed feasible and that studies in non conventional areas for Information Security are the key for developing the proposed goal.

This paper is organized as follows: section 2 introduces some basic definitions concerning automation based attacks that will be used within this paper; section 3 talks about the history of automated human verification and presents the CAPTCHA paradigm; section 4 points out the problems of current CAPTCHA research regarding and presents the problem of a CAPTCHA in the text domain; section 5 outlines the core of our proposal and presents our prototype; section 6 explains the experiment performed with the prototype and discusses our findings; and finally in section 7 some conclusion are drawn.

2 Automated Attacks

Automation based attacks are those which do not actually violate a specific security rule; they simply use (or misuse) a legitimate system in a super-human way, performing requests to the system repeatedly and in a high rate, in order to achieve some objective that goes against the initial goals of the system's designer. Some computer systems are designed with the assumption that only humans will use them, factoring human limitations as a part of the system's security policy. This way, such computer systems do not worry about imposing restrictions on, for example, an excessive high rate of requests per second, because this is simply not humanly possible. The problem happens when super-humans, or computer software robots, enter in action. They can repeat the same request several times every second, disrupting the original intent of the vulnerable computer system. Thus, a simple electronic mailing server, while being under an automation based attack, may become a disgusting SPAM relay.

Stopping automation based attacks within the Internet is a growing trend. This effort has been done basically within two fronts. One approach tries to defeat the mechanisms and techniques that originate the attack, in attempt to stop its very source. An example to that would be a system that, as soon as a super-human use (or automated attack) is identified, it would deny service to the source of the use. The

problem with that approach is that Internet access techniques such as NAT and proxy servers are becoming quite common. Thus, legitimate human users can suffer service denial, as they can easily be misrepresented as attackers. This approach also fails for distributed attacks, which make use of several sources, real and/or fake, such as Distributed Denial of Service (DDoS). It is currently hard, if not unfeasible, to determine the source of such attacks, and service denial to human users would be a problem again. Another problem regarding this approach exists in SPAM filtering, where legitimate emails are sometimes labeled as SPAM. The other approach to counter automated attack works by changing the design of vulnerable systems in a way that potential misuse by automation based attackers is acknowledged as part of the design. This approach is fulfilled with proposals that have as main idea to identify whether or not the entity that is using the system is an authorized party, which means being human. Since systems that are vulnerable to automation based attacks are so due to the fact that they rely on humanity as a characteristic of their users, a protection technique of this sort would have to concentrate on pointing out non-human users, or computational robots, and deny service to them. If a system can be certain that it is being used by human users, high usage, which in other situations would be considered an attack, would not constitute a problem; and if the system knows it is being used by a computer, even during low usage, service denial would take place. Thus, the key to stop automation based attacks is the proper identification of humanity.

3 Turing Tests

Identifying humanity is a complex and long-lasting task. It dates back from the beginnings of modern computer science when Alan Turing presented his theories on the possibility of thinking machines in his famous article “Computing Machinery and Intelligence” [8]. There, Turing proposes his so-called “Imitation Game” (later known as Turing Test), where a human individual would have to interrogate two hidden entities and try to discover their nature concerning humanity. One of the entities would be a human being and the other would be a computer program. Through a series of indirect questions using a computer interface, the human interrogator would have to determine which one was each. This way, the Turing Test was the first test intended to identify humanity within a computational environment.

3.1 Human in the Loop

In 1996, based on Turing’s ideas, Moni Naor proposed a theoretical framework that would serve as the first approach in testing humanity by automated means [11]. In Naor’s humanity test, the human interrogator from the original Turing Test was substituted by a computer program. The original goal of his proposal was to present a scheme that would discourage computer software robots from misusing services originally intended to be used by humans only, much in the same sense of stopping an automation based attack through human identification. Basically, he proposed an adaptation of the way identification is handled in cryptographic settings to deal with this situation. There, when one party A wants to prove its identity to another party B, the process is a proof that A can effectively compute a (keyed) function that a

different user (not having the key) cannot compute. The identification process consists of a challenge selected by B and the response computed by A. What would replace the keyed cryptographic function in the proposed setting would be a task where humans excel, but machines have a hard-time competing with the performance of a three-years-old. By successfully performing such task the user proves that he or she is human.

3.2 CAPTCHAS

Later, Naor's ideas were the basis for a more complete and thorough work on the subject of automated Turing Tests, called then the CAPTCHA paradigm, which was a successful formalization and substantiation of Naor's conceptual model, done by Luis von Ahn et al [1], known as CAPCTHA.

CAPTCHA stands for Completely Automated and Public Turing Test to Tell Computers and Humans Apart. Even though the name itself is self explanatory, some remarks are yet necessary.

Besides formalizing Naor's ideas, von Ahn's work discriminated the important characteristics of an automated Turing Test, leaving aside some of the original concepts that were unnecessary.

As Hard AI problems used by CAPTCHA systems must also be easy for humans to solve, they are generally related to aspects of human cognition. Examples of such problems are optical character recognition (OCR), audio recognition, natural language processing, and image recognition. The same problems for a human being would be, respectively, reading text in images, listening to text in audio samples, understanding the meaning of a text excerpt, and understanding and/or identifying an image sample. It is evident that a human being would have no problems solving those problems, as for a computer program this would not be a trivial task.

Even posing as a difficult task, attackers and security analysts are always trying to find new forms of breaking CAPTCHAs. Be it for self protection or malicious reasons, CAPTCHA systems are constantly subject of attacks and studies that aim to disrupt their efficacy [5,14]. As all the strength of CAPTCHA systems is dependent only on Hard AI problems, breaking a CAPTCHA, in a final analysis, would mean pushing the AI community solving capabilities further ahead. Much in the same sense Naor foretold, the attacker-protector model which is very common in the Information Security field works for CAPTCHAs as a win-win situation, where breaking the system does not only imply a system weakness, but contributes with computer science as a whole improving techniques from other fields.

Therefore, because of its strong formal foundations, the CAPTCHA scheme is the leading research paradigm on automated Turing Tests.

4 Trouble in Paradise

Even though the CAPTCHA framework has a strong formalization and several empirical evidences, CAPTCHA implementations are being extensively broken [5,13,14,16]. Some part of this phenomena falls into von Ahn's objective of

improving AI, but a considerable part is simply a direct result of the exhaustion of CAPTCHA's graphical based model.

Instead of searching for alternate means to explore human cognition, CAPTCHA researchers focus only on basic human senses, such as hearing and specially vision, mainly because they are simple and well known. This tends to push CAPTCHA research onto proposals that mostly try to explore graphical tests, while other possibilities remain open problems, such as the problem of a CAPTCHAs in the text domain.

A CAPTCHA in the text domain (or text based CAPTCHA) would mainly explore linguistic cognition aspects of humans, or the ability humans have to understand linguistic constructs. The construction of a CAPTCHA in the text domain is often cited as an important open problem [1,4,6]. To our knowledge, the only formal attempts to construct a CAPTCHA of this sort are [4] and [15], but they all fail to address the issue.

In [4], a word from a piece of text taken from a data source of human-written text is randomly selected and substituted by another word selected at random, in the hope that it would be easy for humans to pick that word (because it didn't fit in the context), but difficult for computers. However, it is demonstrated also in [4] that it was possible to write a program that had considerable success-rates in "cheating" the test by taking into account statistical characteristics of natural language.

In [15], it is proposed the use of lexical semantics to construct an HIP that draws its security from the problem of word-sense ambiguity, i.e., the phenomenon that a single word can have different meanings and that different words can have the same meaning, depending on the context in which a word is used. Despite the fact that indeed this HIP proposes a task difficult for computers and easy for humans, it violates Kerckhoff's principle [12] that is present in the CAPTCHA paradigm, as the efficacy of the test is based on the secrecy of the database that holds the "secret annotations", which are mappings necessary for the disambiguation process, which is all it is necessary to solve the test. Furthermore, it is not very clear how this database would be constructed and the author only indicates that it is necessarily constructed with human intervention possibly creating a barrier for automating the test.

5 A CAPTCHA in the Text Domain

5.1 Proposal

Through a general overview of some possibilities for the deployment of our CAPTCHA, we decided to concentrate on a particular work by Julian Taylor [9]. She has studied automated (computational) generation and recognition of humorous constructs on the focused domain of Knock-Knock (KK) jokes. A KK joke is basically a type of humorous punning (wordplay) riddle. A regular KK joke is a dialog between two people that uses wordplay in the punch line. A KK joke can be summarized using the following structure:

Line1: "Knock, Knock"

Line2: "Who is there?"

Line3: any phrase

Line4: Line3 followed by “who?”

Line5: One or several sentences containing one of the following:

Type1: Line3

Type2: a wordplay on Line3

Type3: a meaningful response to Line3.

KK jokes are more common in the English speaking world, though its structure provided us with important characteristics that we believe may assist in our goals. These are:

1. A KK joke is a linguistic construct that by its humorous nature becomes easily recognizable and sometimes enjoyable.

2. Despite the fact KK jokes are not cultural available worldwide, one may argue that phonetic punning riddles are.

3. A regular KK joke is a simple and stable structure, with a formation rule.

4. KK jokes are based on phonetic punning riddles as they explore cognitive aspects not only related to linguistics, but also to sound interpretation. This way, we empower ourselves with more tools in order to build the proposed CAPTCHA.

5. There is evidence of an incongruity between computation KK joke generation and computational KK joke understanding. This conclusion was drawn by some of the remarks found on Taylor’s work on KK Jokes. She was able to build a successful KK joke generator, but was not capable of building a KK joke recognizer that could in fact do its job. Despite her efforts, Taylor’s KK Joke recognizer could only find wordplays, but was unable to determine if the joke made sense or not. She justifies that by stating that the creation of KK jokes requires restrict knowledge, whereas their understanding requires “world” knowledge. We believe this gap is sufficient enough to generate humorous text excerpts based on the KK joke structure that computers will not “get”.

Our proposed scheme consists basically of a challenge that would present a set of KK Joke like structures to the user. Despite all of the presented structures would be built upon the same general linguistic structure, only one of them would make sense as a real KK joke. The user would have to indentify the correct joke within the set in order to prove his human condition. Therefore, our proposed CAPTCHA concentrates

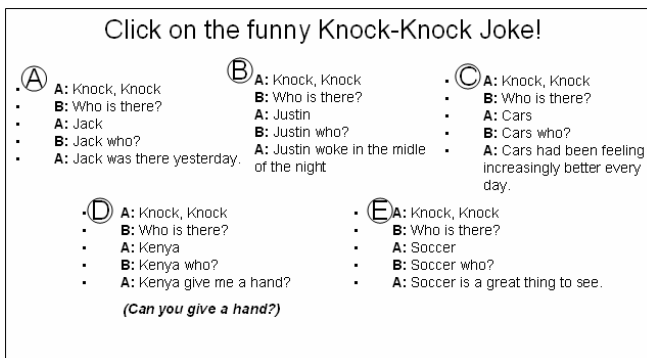


Fig. 1. Example of the proposed scheme

on a test that would generate fake jokes following the same structure of a regular KK Joke (or a variation of it), together with a real KK joke. By correctly identifying the real KK Joke, a human individual would be set apart from computational robots. Figure 1 shows an example of the proposed CAPTCHA, for the English language.

5.2 The Prototype

In order to understand human cognitive impacts of an automatically generated KK Joke and to create the foundations for and at the same time improve our model, we have developed a basic prototype for our CAPTCHA system.

The prototype consisted basically of a double challenge test, where a user was required to pin point the non-bogus KK joke among a given set. Each challenge presents one real KK joke and two fake KK jokes. By successfully identifying the non-fake KK joke on each one of the two challenges, the user proves that he or she is human.

To strengthen our contend that internationalization would not be an issue to the development of our proposal and for the obvious reason of being mostly a Brazilian research team, we have developed our prototype in Portuguese. Therefore, we have coined a variation of a regular KK joke, since there is no such structure in the Portuguese language. It is basically a simpler KK joke based structure in Portuguese, as follows:

Line1: “Você conhece <wordplay>?”

Line2: “<wordplay> quem?”

Line3: a meaningful response starting by <wordplay>

The translation for that would be:

Line1: “Do you know <wordplay>?”

Line2: “<wordplay> who?”

Line3: a meaningful response starting by <wordplay>

The reason we did not performed a literal translation of the same structure found on a regular KK Joje is that we believe the KK joke structure is too long. This doe not constitute a problem for native speakers of English since cultural background helps in

- Select a random sentence;
- Make a phonetic transformation on the first two words from the random sentence;
- Turn the phonetic transformation into a single word, or the phonetic transformation word (PTW).
- Substitute the first two words of the random sentence by the PTW, in order to create the transformed sentence (TS).
- Use PTW and TS to build the following structure:
 - Knock Knock!
 - Who's there?
 - PTW.
 - PTW, who?
 - TS

Fig. 2. Real Joke Generation Algorithm

attenuating this problem. Since our translation could not count on any cultural background from the users it had to be simpler.

The KK joke generation process explored phonetic puns (wordplays) as a means of creating meaningful jokes. The generation algorithm of a real KK joke is summarized by the steps in figure 2.

On the other hand, the generation algorithm for a fake KK joke is summarized in figure 3.

- Generate two non-fake KK jokes;
- Exchange their PTWs;
- Randomly pick one of changed jokes;
- Exclude the other;

Fig. 3. Fake Joke Generation Algorithm

The phonetic transformation process used in both algorithms followed some of the basic ideas proposed by Taylor. Basically it consists of a process that changes consonantal phonemes by similar ones, based on the phoneme similarity table proposed by [10], that can be found in figure 4.

The resulting structure is a sentence that resembles its original version, mainly because sound similarities. As our prototype was designed for Portuguese language, we have adapted it to use the phoneme similarity table in such a way that only phonemes found on Portuguese language would be considered.

	p	b	f	v	m	t	d	θ	ð	s	z	ʃ	ʒ	tʃ	dʒ	k	g	ŋ	l	r	n	w	y	h
p	1																							
b	0.4	1																						
f	0.26	0.13	1																					
v	0.15	0.3	0.38	1																				
m	0.19	0.39	0.07	0.15	1																			
t	0.3	0.14	0.1	0.06	0.06	1																		
d	0.14	0.28	0.05	0.11	0.11	0.39	1																	
θ	0.11	0.06	0.43	0.19	0.03	0.2	0.11	1																
ð	0.07	0.12	0.19	0.39	0.06	0.12	0.23	0.38	1															
s	0.1	0.05	0.18	0.1	0.03	0.3	0.15	0.4	0.2	1														
z	0.06	0.11	0.09	0.19	0.06	0.17	0.33	0.19	0.44	0.37	1													
ʃ	0.1	0.05	0.18	0.1	0.03	0.18	0.1	0.4	0.2	0.58	0.24	1												
ʒ	0.06	0.11	0.09	0.19	0.06	0.11	0.2	0.19	0.44	0.24	0.57	0.37	1											
tʃ	0.21	0.11	0.1	0.06	0.06	0.44	0.22	0.21	0.13	0.27	0.14	0.41	0.21	1										
dʒ	0.11	0.22	0.06	0.11	0.11	0.22	0.47	0.11	0.24	0.13	0.28	0.19	0.44	0.39	1									
k	0.44	0.19	0.14	0.08	0.08	0.35	0.16	0.13	0.08	0.11	0.06	0.11	0.06	0.25	0.13	1								
g	0.21	0.42	0.08	0.16	0.15	0.17	0.33	0.07	0.15	0.06	0.13	0.06	0.13	0.14	0.27	0.39	1							
ŋ	0.09	0.15	0.04	0.09	0.37	0.07	0.13	0.04	0.08	0.04	0.07	0.04	0.07	0.13	0.17	0.33	1							
l	0.04	0.07	0.04	0.08	0.17	0.11	0.19	0.08	0.17	0.11	0.22	0.07	0.14	0.07	0.13	0.05	0.09	0.24	1					
r	0.1	0.19	0.07	0.14	0.44	0.09	0.16	0.06	0.13	0.09	0.18	0.06	0.11	0.06	0.11	0.04	0.07	0.17	0.56	1				
n	0.06	0.12	0.03	0.06	0.26	0.19	0.38	0.06	0.13	0.09	0.18	0.06	0.11	0.12	0.24	0.07	0.14	0.33	0.53	0.4	1			
w	0.14	0.25	0.09	0.19	0.44	0.03	0.06	0.04	0.08	0.04	0.07	0.04	0.07	0.04	0.06	0.05	0.09	0.18	0.17	0.42	0.12	1		
y	0.04	0.07	0.04	0.09	0.13	0.07	0.13	0.08	0.17	0.07	0.14	0.12	0.23	0.12	0.21	0.05	0.09	0.18	0.40	0.29	0.27	0.25	1	
h	0.15	0.08	0.47	0.21	0.04	0.12	0.06	0.41	0.19	0.23	0.11	0.23	0.11	0.13	0.07	0.19	0.1	0.06	0.06	0.04	0.04	0.06	0.06	1

Fig. 4. Phoneme Similarity Table

6 Experiment

6.1 Trial

In order to experiment with our CAPTCHA prototype, we have developed a free SMS messaging WEB site that used our prototype. The site offered free SMS messaging to the main Brazilian cell phone operators, including one that normally charges for this service, even via web. We believe this was just enough incentive in a way that would not compel users to forcedly use our system.

We have run the prototype for two days. During this period, a total of 584 users came to have contact with the system, but only 455 of them actually tried to use it at least once. During the experiment a total of 894 tests were performed. Taking into consideration that our prototype presented a double challenge test, another form of analysis would be to consider each challenge alone. The total of single challenges was 1893. A total of 221 tests were answered correctly, which equals 24.72% of the total of answered tests. Analyzing by the challenges point of view, a total of 887 were answered correctly (46.86% of the amount of answered challenges). At first this seems a little discouraging, but after further analysis we noticed that some challenges were just answered too fast, some even in approximately 0 seconds. If one considers that each challenge is presented with three text excerpts, it is possible to argue that a fast glimpse at each one of them would require at least 2 or 3 seconds and the whole challenge would require at least 6 seconds to be answered. We believe this anomaly happened due to some attempts in using a computational robot to break our system (which we further confirmed to be true).

This way, we decided to filter the results by answering time. This led to new and less discouraging results which we summarized in table 1 for complete tests and table 2 for challenges alone.

Table 1. Success percentages for complete tests

Minimum Time Spent	Total of Answered Tests	Total of Correct Answers	Success Percentage
0 seconds	894	221	24,72%
5 seconds	550	181	32,9%
15 seconds	443	147	33,18 %

Table 2. Success percentages for single challenges

Minimum Time Spent	Total of Answered Tests	Total of Correct Answers	Success Percentage
0 seconds	1893	887	46,86%
5 seconds	1258	677	53,81%
15 seconds	1112	602	54,11 %

7 Conclusions

This paper has shown an innovative approach in CAPTCHA research by presenting a strong proposal for a CAPTCHA in the text domain. Though yet inconclusive, our results indicate that some generated KK jokes share some particular characteristics that permit humans to point them out as real jokes. A random guess in our experiment would generate the probability of 11,11% of success for a complete test and 33,33% of success for challenges alone. Taking into consideration that the best results for challenges were 54,11% of success, there is an advantage 20,78% of success over random chance. If we analyze complete test this gap is even bigger, equaling 22,07% (33,18% - 11,11%). This is specially encouraging, mainly because our prototype is yet in its first version where several issues may still be perfected and new concepts are yet to be incorporated. Taking all these factors into consideration, our findings indicate a real feasibility of building a CAPTCHA of the proposed sort.

Acknowledgments. This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 200503311809.

References

1. Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: using hard ai problems for security. In *Advances in Cryptology, Eurocrypt 2003*, volume 2656 of Springer Lecture Notes in Computer Science, pages 294–311, May 2003.
2. Tsz-Yan Chan. Using a text-to-speech synthesizer to generate a reverse turing test. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, page 226. IEEE Computer Society, 2003.
3. Allison L. Coates and Richard J. Fateman. Pessimial print: A reverse turing test. In *Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, 2001.
4. Philip Brighten Godfrey. Text-based CAPTCHA algorithms. In *First Workshop on Human Interactive Proofs*, 2002. Unpublished Manuscript. Available electronically: http://www.aladdin.cs.cmu.edu/hips/events/abs/godfreyb_abstract.pdf.
5. Greg Mori and Jitendra Malik. Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA. In *Conference on Computer Vision and Pattern Recognition (CVPR '03)*, volume I, 2003.
6. Bartosz Przydatek. On the (im)possibility of a text-only CAPCHA. In *First Workshop on Human Interactive Proofs*, 2002. Unpublished Abstract. Available electronically: http://www.aladdin.cs.cmu.edu/hips/events/abs/bartosz_abstract.pdf.
7. Graeme Ritchie “Prospects for Computational Humour,” *Proceedings of 7th IEEE International Workshop on Robot and Human Communication*, Takamatsu, Japan, pp. 283-291, 1998
8. Alan M. Turing. Computing machinery and intelligence. *Mind*, 49:433–460, 1950. 19. Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. Hips. <http://www.aladdin.cs.cmu.edu/hips/>.
9. Julia Taylor, “Computational Recognition of Humor in a Focused Domain”, Master Thesis, University of Cincinnati, 2004.
10. Stefan Frisch, “Similarity And Frequency In Phonology”, Doctoral dissertation, Northwestern University, 1996

11. Moni Naor. Veri_cation of a human in the loop or Identi_cation via the Turing Test. Unpublished Manuscript, 1997. Available electronically: <http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.ps>.
12. Auguste Kerckhoffs, La cryptographie militaire, Journal des sciences militaires, vol. IX, pp. 5–83, Jan. 1883, pp. 161–191, Feb. 1883.
13. Chellapilla K., and Simard P., “Using Machine Learning to Break Visual Human Interaction Proofs (HIPs),” Advances in Neural Information Processing Systems 17, Neural Information Processing Systems, MIT Press, 2004
14. Gabriel Moy, Nathan Jones, Curt Harkless, and Randall Potter Distortion Estimation Techniques in Solving Visual CAPTCHAs Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2004
15. Richard Bergmair and Stefan Katzenbeisser, “Towards Human Interactive Proofs in the Text-Domain Using the Problem of Sense-Ambiguity for Security” 7th International Conference, ISC 2004, Palo Alto, CA, USA, September 27-29, 2004
16. The OCR Reaearch Team, “Weak CAPTCHAs”, 2006 available online in: <http://ocr-research.org.ua/list.html>

Examining the DoS Resistance of HIP

Suratose Tritilanunt, Colin Boyd,
Ernest Foo, and Juan Manuel González Nieto

Information Security Institute
Queensland University of Technology
GPO Box 2434, Brisbane, QLD 4001, Australia
s.tritilanunt@student.qut.edu.au,
{c.boyd, e.foo, j.gonzaleznieto}@qut.edu.au

Abstract. We examine DoS resistance of the Host Identity Protocol (HIP) and discuss a technique to deny legitimate services. To demonstrate the experiment, we implement a formal model of HIP based on Timed Petri Nets and use a simulation approach provided in CPN Tools to achieve a formal analysis. By integrating adjustable puzzle difficulty, HIP can mitigate the effect of DoS attacks. However, the inability to protect against coordinated adversaries on a hash-based puzzle causes the responder to be susceptible to DoS attacks at the identity verification phase. As a result, we propose an enhanced approach by employing a time-lock puzzle instead of a hash-based scheme. Once the time-lock puzzle is adopted, the effect of coordinated attacks will be removed and the throughput from legitimate users will return to the desirable level.

1 Introduction

Many key exchange protocols have been developed for dealing with denial-of-service (DoS) attacks, especially resource exhaustion attacks. Host Identity Protocol (HIP) [14] is an interesting example of a DoS-resistant protocol which has been developed to deal with this kind of DoS attack. The concept behind this implementation is that HIP does not commit the responder's resource before the responder ensures the identity of the initiator. HIP achieves this concept by adopting stateless connection [3] and reachability testing by using a *client puzzle* [4, 12] incorporated via a cookie [16] to protect the responder from SYN flooding attacks [7] at the beginning phase. Moreover, the responder can authenticate the initiator by starting with the cheap computation using a client puzzle and then increase the level of authentication to the expensive computation using a digital signature for ensuring the identity of the initiator.

HIP is a promising key exchange protocol which includes DoS-resistant mechanisms for protecting the responder. However, lack of formal analysis in the design phase of HIP might introduce other kinds of vulnerability. Moreover, the instruction on how to adjust the client puzzle difficulty is not clearly specified and examined in the HIP specification [14]. In this paper, we implement a formal model of HIP using the formal specification language of Timed Petri Nets.

In order to achieve a formal analysis, we use a simulation technique provided in CPN Tools for analysing HIP model. The purpose of the simulation in the cryptographic protocol is to identify vulnerabilities in the system that might be difficult to explore in the design phase.

Simulation approaches are well-known not only for exploring vulnerabilities in cryptographic protocols, but guaranteeing security services of such protocols as well. Using simulation approaches has several benefits over mathematical analysis. For instance, they can provide *flexibility* and *visualization* during protocol analysis and verification. In our experiment, we set up the simulation of HIP for exploring unbalanced computational steps that cause a responder to spend more computations than an initiator does. In addition, our experimental result provides a measurement of successful legitimate traffic as proposed by Beal and Shepard [6] in different situations under DoS attacks. This factor can be used as a parameter for justifying the effectiveness of HIP to resist DoS attacks. In order to set up an experiment, we allow four kinds of adversary and the honest client to participate with the same responder during the protocol run. We set up two experiments; 1) the responder can choose only a fixed value of a puzzle difficulty no matter what the workload is, and 2) the responder has an ability to flexibly adjust puzzle difficulty by using the workload condition as criterion.

The main contributions of this paper are:

1. A simulation and analysis of HIP in Timed Coloured Petri Nets.
2. Identification of four scenarios of resource exhaustion attack on HIP.
3. A proposed technique to deal with adversaries who try to overwhelm the responder's resource by computing a puzzle solution in parallel.

1.1 Host Identity Protocol (HIP)

HIP has been developed by Moskowitz [14]. Later, Aura et al. [2] found some vulnerabilities and proposed guidelines to strengthen its security. HIP is a four-packet exchange protocol which allows the initiator I and responder R to establish an authenticated communication. Both I and R hold long-term keys to generate signatures $Sig_I(\cdot)$ and $Sig_R(\cdot)$ respectively. It is assumed that both principals know the public key PK_I of the initiator and PK_R of the responder represented in the form of host identifiers (HI) in advance. HIT represents the host identity tag created by taking a cryptographic hash H over a host identifier.

H_{K_s} represents a keyed hash function using session key K_s to generate a hashed-MAC ($HMAC$). The value s is a periodically changing secret only known to the responder. LSB takes as input a string t and a parameter k and returns the k least significant bits of t . 0^k is a string consisting of k zero bits. $E_{K_e}(\cdot)$ and $D_{K_e}(\cdot)$ denotes a symmetric encryption and decryption respectively under session key K_e . In order to generate session keys K_e and K_s , HIP employs Diffie-Hellman key agreement. Diffie-Hellman parameters used to generate these keys consist of large prime numbers p and q , a generator g , a responder's secret value r , and an initiator's secret value i .

HIP adopts a proof-of-work scheme [11] for countering resource exhaustion attacks. In a proof-of-work, HIP extends the concept of a *client puzzle* [4, 12]

	<i>I</i>	<i>R</i>
		<i>Precomputed parameters</i>
		$r, s \in_R [1, 2, \dots, q - 2]$
		$sig_{R1} = Sig_R(g^r, HIT_R)$
	
1) create HIT_I, HIT_R	$\xrightarrow{HIT_I, HIT_R}$	check HIT_R
		$C = LSB(H(s, HIT_I, HIT_R), 64)$
2) verify sig_{R1}	$\xleftarrow{HIT_I, HIT_R, puzzle, g^r, sig_{R1}}$	$k \in [0, 1, \dots, 40] \rightarrow puzzle = (C, k)$
Find J such that		
$LSB(H(C, HIT_I, HIT_R, J), k) = 0^k$		
$i \in_R [1, 2, \dots, q - 2]$		
$K_e = H(HIT_I, HIT_R, g^{ir}, 01)$		
$E1 = E_{K_e}\{HI_I\}$		
$sig_I = Sig_I(HIT_I, HIT_R, J, g^i, E1)$		
3)	$\xrightarrow{HIT_I, HIT_R, J, g^i, E1, sig_I}$	$C = LSB(H(s, HIT_I, HIT_R), 64)$
		$LSB(H(C, HIT_I, HIT_R, J), k) \stackrel{?}{=} 0^k$
		$K_e = H(HIT_I, HIT_R, g^{ir}, 01)$
		decrypt $E1$
		verify sig_I
		$K_s = H(HIT_I, HIT_R, g^{ir}, 02)$
		$HMAC = H_{K_s}(HIT_I, HIT_R)$
4) verify sig_{R2}	$\xleftarrow{HMAC, sig_{R2}}$	$sig_{R2} = Sig_R(HIT_I, HIT_R, HMAC)$
$K_s = H(HIT_I, HIT_R, g^{ir}, 02)$		
$H_{K_s}(HIT_I, HIT_R) \stackrel{?}{=} HMAC$		

Fig. 1. HIP Protocol [14]

for protecting the responder against DoS attacks. HIP uses the client puzzle to delay state creation [3] in the responder until the checking of the second incoming message and the authentication has been done in order to protect the responder against resource exhaustion attacks.

1.2 Previous Work

Over many years, cryptographic and security protocols have been modeled and verified using Coloured Petri Nets (CPNs). Doyle [8] developed a model of three-pass mutual authentication and allowed an adversary to launch multiple iteration and parallel session attacks. Han [10] adopted CPNs for constructing a reachability graph to insecure states and examining the final states in OAKLEY.

Al-Azzoni [1] developed a model of Needham-Schroeder public-key authentication protocol and Tatebayashi-Matsuzaki-Neuman (TMN) key exchange protocol.

Beal and Shepard [6] constructed a model of HIP protocol using a mathematical equation for analysing the effect of puzzle difficulty under the steady-state attack. In order to deamplify the arrival rate of incoming requests, Beal and Shepard have set up two strategies; 1) forcing a sustainable arrival rate, and 2) limiting service disruption. They modeled adversaries with the ability of adversaries has been limited to disrupt the service of legitimate initiators by flooding bogus requests.

To the best of our knowledge, there is no implementation of CPNs focusing on an exploration of vulnerabilities based on unbalanced computation that might lead to resource exhaustion attacks in key exchange protocols. Moreover, Beal and Shepard's mathematical model has a few limitations including 1) they do not allow the responder to dynamically adjust puzzle difficulty, and 2) there is only one attacking technique to overwhelm the responder's resources.

2 Experimental Results and Analysis

In our model, we have allowed a system to consist of honest clients, individual type of adversaries, and a responder. The responder has to deal with different strategies of adversaries and amounts of packets which consist of both legitimate and bogus messages. We allow three different packet rates for both honest clients and adversaries in order to measure the toleration of HIP under DoS attacks. Honest clients can initiate the amount of requests (C) at 80%, 100%, and 150% of the responder's capacity (R). Meanwhile, a single type of adversary can flood the amount of bogus requests (Z) at 100%, 200%, and 1000% of the responder's capacity (R).

Apart from honest clients (hc) who initiate the legitimate traffic, we allow four types of adversary who have the similar goal to deny the service of the responder by overwhelming CPU usage and connection queue of the responder. While other adversarial strategies are certainly possible, the defined adversaries cover the most obvious attacks at all stages of the protocol execution. To our knowledge, no previous formal analysis of DoS-resistant protocols has included such a comprehensive adversary definition.

Type 1 adversary (ad1) computes a valid first message (may be pre-computed in practice), and then takes no further action in the protocol.

Type 2 adversary (ad2) completes the protocol normally until the third message is sent and takes no further action after this. The computations of this adversary include searching a correct client puzzle solution J , generating a session key K_e and encrypting a public key PK_I , and finally computing a digital signature Sig_I .

Type 3 adversary (ad3) completes the protocol step one and two with the exception that the adversary does not verify the responder signature sig_{R1} .

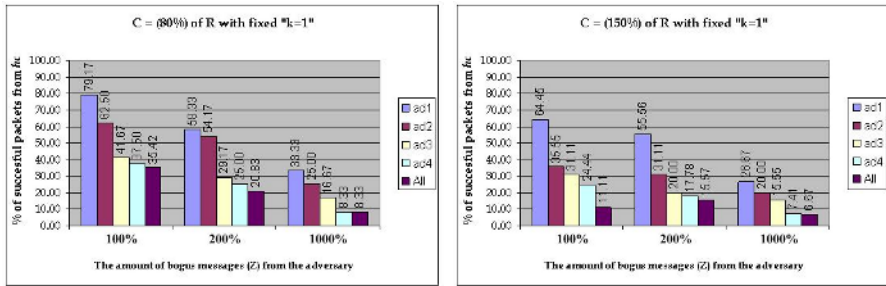
The adversary searches for a correct client puzzle solution J but randomly chooses the remaining message elements: an encrypted element $K_e\{HI_I\}$ and a digital signature sig_I . The adversary takes no further action in the protocol.

Type 4 adversary (ad4) is like an adversary type 3, except that the client puzzle solution J is now also chosen randomly.

In the simulation, we initially set up the responder’s capacity for handling incoming requests. The **hc** initiates a request only once and keeps waiting to process next steps. If its request is rejected, **hc** gives up. For adversaries, there are two different situations in which the responder rejects bogus messages; 1) the responder detects the bogus messages during the verification steps, and 2) the responder does not have enough resources for serving requests. In order to evaluate the system performance, the rate of successful legitimate traffic under different attacks has been measured as the percentage of throughput. Some sample results of our experiment are demonstrated in the following subsection.

Experiment 1: Non-adjustable client puzzle

The purpose of experiment 1 is to examine the minimal DoS-resistant mechanism. To achieve this, we run the simulation under four specified attacks and the combination of four strategies (defined as *All*) with the non-adjustable client puzzle. We initially fix $k=1$, i.e. the easiest value¹, because **hc** prefers to spend nothing expensive for establishing a connection under normal circumstances.



(a) $hc = 80\%$ (b) $hc = 150\%$

Fig. 2. Percentage of throughput from honest clients with $k=1$

Figure 2 represents the percentage of successful legitimate connections compared among three different amount of bogus messages ($Z=100\%$, 200% , and 1000% of the responder’s capacity R) from five adversarial strategies (in the combination strategy, *All*, each of adversary type has the same amount of bogus messages that makes the total number equivalent to the specified quantity).

From figure 2, when adversaries increase the number of bogus messages, the percentage of successful messages from **hc** to obtain a service will drop drastically. Comparing **ad1** and **ad4**, even though both of them craft random messages,

¹ If we choose $k=0$, we cannot see the difference of costs between **ad3** and **ad4**.

ad4 can achieve the goal at higher rate than ad1 because the responder can process the incoming request at step 1 and clear a queue faster than at step 3. At step 1, the responder only participates in the protocol by choosing the puzzle difficulty (k) and pre-computed information, and returns it to ad1. Although, ad1 can re-send bogus messages after receiving replied messages, this does not cause the responder to reject a large number of messages because HIP mitigates such problem by adopting a stateless-connection. On the other hand, the task of ad4, to fill-up the responder's queue at step 3, can be achieved more easily than ad1 because the process of checking a puzzle solution and a digital signature takes longer than a whole process at step 1.

Comparing ad2 and ad3 who attempt to deny service at phase 3 by computing the puzzle solution, the results show that ad3 succeeds at higher proportion than ad2. This is because ad3 can flood attack messages faster than ad2 who must engage in the correct generation of message two. Nonetheless, both adversaries can force the responder to engage in the signature verification. Although ad4 can flood large number of messages at step 3 as well as ad2 and ad3, ad4 cannot force the responder to engage in expensive operations because the responder is able to detect the message forgery at the cheap puzzle verification process. However, without the assistance of puzzle difficulty, the percentage of successful messages in the case of hc and ad4 is lower than the others because ad4 floods message three at the highest rate. As a result, the most effective adversary to deny services on the responder would be ad4 that attacks the verification phase. Most key agreement protocols incorporate verification tasks that would be susceptible to resource exhaustion attacks.

The result of the combination of all attack techniques shows that when the responder has to deal with all types of adversary, the percentage of legitimate users served by the responder will fall significantly with increase of bogus messages. Now we have identified the most effective scenario, we will apply this technique to the experiment 2 for investigating the usefulness of puzzle difficulty.

Experiment 2: Adjustable client puzzle

The purpose of the second experiment is to observe and evaluate how a client puzzle can mitigate the problem of DoS attacks on the responder's machine. By calibrating several ranges of puzzle difficulty to obtain an optimal throughput, we anticipate to find a simple and flexible technique for dynamically adjusting puzzle difficulty to suit all DoS-attack scenarios.

To adjust the puzzle difficulty, we allocate two possible values for the responder to determine. Under normal circumstances, the responder selects $k=1$, which means the easiest puzzle solution is required from the initiator. Once the responder receives more requested packets than its maximum capacity to handle, the responder raises the puzzle difficulty. In the experiments described here, we choose $k=10$. Because this puzzle technique is a hash-based puzzle, this value will help the responder to slow down the incoming rate by requiring the work of the initiator to solve a puzzles at the factor of 2^{10} .

Similarly to the representation of Figure 2, Figure 3 illustrates that the number of attacking machines that the responder can tolerate is increased to a higher

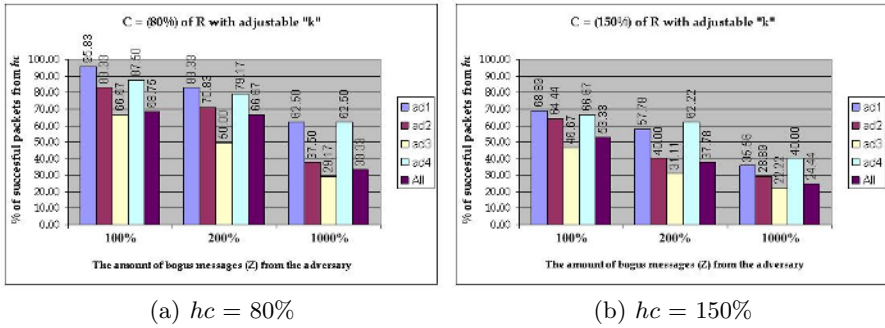


Fig. 3. Percentage of throughput from hc with k is adjustable between 1 and 10

proportion compared to the result of experiment 1. Another interesting result is that the successful rate of an honest client’s message in the case of ad4 is higher than for the fixed value $k=1$. The reason is that ad4 does not compute the puzzle solution, so, no matter what the puzzle difficulty is, ad4 can flood the bogus messages at the similar speed as experiment 1. However, at that amount of bogus messages, there are only messages from ad4 (no legitimate traffic because hc has to spend some amount of time to solve the puzzle solution), or just a few messages from hc that arrive to the connection queue before the responder increases puzzle difficulty. As a result, the responder can validate the puzzle solution before the next group of messages has arrived. Undoubtedly, these bogus messages from ad4 will be rejected at the first step of verification which requires only short period and removes such attack from the connection queue. However, this situation does not occur in the case of ad3 because they have to spend some amount of time to solve the puzzle as well as hc.

In experiment 2, the most effective scenario is from ad3. Comparing to experiment 1, if adversaries can flood messages at the same speed as ad4 and force the responder to participate in expensive verification as for ad3, those adversaries would obtain higher satisfied outcome. A possible adversarial technique to obtain higher rejected rate is that if ad3 can solve a puzzle more quickly and flood these solutions as fast as ad4. These packets will be accumulated in the connection queue longer than those from ad4 because the responder has to participate and spend more computational time to verify ad3’s messages due to the signature verification. To achieve this technique, suppose that we use SHA-1 for generating hash output, so the result is 160-bits long. When the responder chooses $k = 10$, it means that the 10 left-most significant bits must be zero but the remaining bits can be either 1 or 0. Therefore, the chance of a user to get the result which has 10-zero bits at the beginning of the output would be 2^{-10} . As a result, if ad3 shares value J , which is 64-bits long, to coordinated attackers (Co-ad) trying to find a solution, they can save time in this process depending on the number of participating machines in Co-ad. They can achieve this technique because a puzzle construction based on hash function can be computed faster

in a parallel fashion. This attack technique is defined in terms of a coordinated attack [17]. The experiment and results of are demonstrated in Section 3.

3 A New Approach

Vulnerabilities based on unbalanced computations between an initiator and a responder have been revealed in Section 2. This vulnerability leads to the risk of the responder’s machine to be overwhelmed by the coordinated adversaries. This section propose a technique to mitigate this problem. The results show that the proposed technique can help to deal with such attack.

In the experiment, we re-construct a HIP model by adopting the concept of a time-lock puzzle [15] which has been developed by Rivest et al. The fundamental property of time-lock puzzles is that they require a precise amount of time to be solved and can not be solved in parallel computation. Therefore, the responder can select the predetermined time period for a puzzle for delaying the incoming requests when the responder has heavy load to serve.

In order to generate a time-lock puzzle, the responder has to determine the amount of time for the client to spend in solving the puzzle (T) and estimate the initiator capacity in calculating repeated squaring per second (S). Next, the responder computes the number of repeated squaring $t = T \cdot S$ that must be computed by the initiator in order to find a solution. Finally, the responder forces the initiator to calculate $b = a^{2^t} \pmod{n}$, where n is the product of two large primes p and q . Because the responder knows the factors p and q , he can compute b much faster by first computing $2^t \pmod{\phi(n)}$.

We setup simulation for evaluating a system corresponding to coordinated adversaries type 3 (**Co-ad3**). When we insert a time-lock puzzle into HIP model at step two of the initiator, the result for **hc** and **Co-ad3** will be improved to the higher percentage of successful packets approximately equal to the experiment 2. In the experimental results, graphs represented with **Co-ad3** term are simulated by using a hash-based puzzle with adjustable k , while graphs represented with **k=1** and **varied-k** are simulated by using a time-lock puzzle with fixed $k=1$, and adjustable k , respectively.

Not similar to the representation of Figures 2 and 3, the x-axis of Figure 4 compares among three defined values of the legitimate requests ($C=100\%$, 80% , and 150% of the responder’s capacity R). From the result, if we compare the graph of **ad3** at workload **hc** = 80% of R in Figure 3(a) with **Co-ad3** in Figure 4(a), the throughput falls from 66.67% to 37.50% . Once we employed time-lock puzzle as shown in the graph **k=1** and **varied-k** of figure 4, the throughput will increase to approximately the same as experiment 2 (shown in figure 3). The reason is that **Co-ad3** has been forced to spend time specified by the responder until the time-lock puzzle has been solved. This period is similar to the period in experiment 2 in which normal **ad3** spends time to search for a correct solution of a hash-based puzzle. As a result, when the responder constructs a time-lock puzzle, the responder can control time required for the initiator to solve a puzzle

more precisely. Figure 4 displays results from the simulation which adopts the time-lock puzzle technique into the system. We see that use of a hash-based puzzle against a coordinated adversary results in less throughput than no puzzle at all ($k=1$). Use of a time-lock puzzle with varied k effectively increases the percentage of successful packets from the hc.

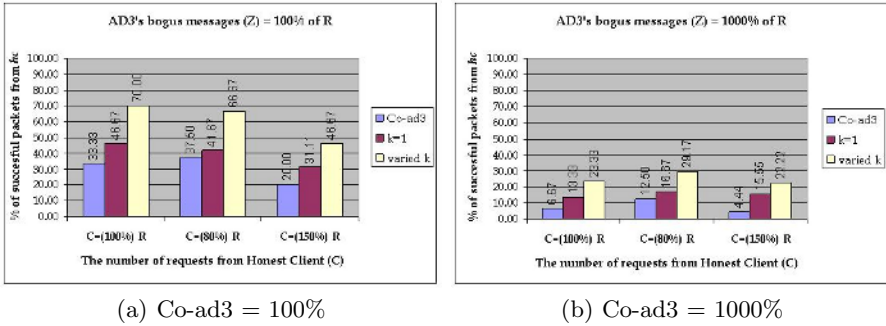


Fig. 4. Percentage of throughput from *hc* and *Co-ad3* comparing between a *hash-based* puzzle and a *time-lock* puzzle

4 Conclusion and Future Work

According to a comparison by Feng et al. [9], the most interesting property of time-lock puzzle is non-parallelizability that prevents *Co-ad* to speed up the process of searching a solution by distributing a puzzle to other high-performance machines. Moreover, time-lock puzzles also provide fine-grained control in order to precisely adjust puzzle difficulty by the responder. Although the integration of time-lock puzzles mitigates the problem of *Co-ad3*, the underlying computation for constructing time-lock puzzle is a major concern because the puzzle generation is limited by the calculation of modular exponentiation which has greater magnitude than hash-based puzzle. Some example uses of time-lock puzzles have been evaluated and identified by Mao [13], Back [5], and Feng et al. [9] which all suffer from the same problem.

It would be useful to discover new techniques to construct a client puzzle satisfying desirable properties identified by Aura et al. [4]. In particular, puzzles should be inexpensive for the responder to generate and verify, and impossible to precompute a solution by the initiator. Two additional properties which enhance DoS-resistant protocols for preventing *Co-ad3* should be included:

1. the puzzle should not be solvable in parallel for obtaining an output in less than a specific time.
2. the responder should be able to precisely control puzzle difficulty in a linear manner.

References

1. I. Al-azzoni. The Verification of Cryptographic Protocols using Coloured Petri Nets. Master of Applied Sciences Thesis, Department of Software Engineering, McMaster University, Ontario, Canada, 2004.
2. T. Aura, A. Nagarajan, and A. Gurtov. Analysis of the HIP Base Exchange Protocol. In *Proceedings of 10th Australasian Conference on Information Security and Privacy (ACISP 2005)*, pages 481 – 493, Brisbane, Australia, Jun 2005.
3. T. Aura and P. Nikander. Stateless Connections. In *International Conference on Information and Communications Security*, pages 87–97, Beijing, China, Nov 1997.
4. T. Aura, P. Nikander, and J. Leiwo. DoS-resistant authentication with client puzzles. In *Security Protocols Workshop 2000*, pages 170–181, Apr 2000.
5. A. Back. Hashcash - A Denial of Service Counter-Measure, 2002. <http://citeseer.ist.psu.edu/back02hashcash.html>.
6. J. Beal and T. Shepard. Deamplification of DoS Attacks via Puzzles. Available: <http://web.mit.edu/jakebeal/www/Unpublished/puzzle.pdf>, 2004.
7. Computer Emergency Response Team (CERT). SYN Flooding Attack. [Online]. Available: <http://www.cert.org/advisories/CA-1996-21.html>, 1996.
8. E. M. Doyle. Automated Security Analysis of Cryptographic Protocols using Coloured Petri Net Specification. Master of Science Thesis, Department of Electrical and Computer Engineering, Queen’s University, Ontario, Canada, 1996.
9. W. Feng, A. Luu, and W. Feng. Scalable, Fine-grained Control of Network Puzzles. Technical report 03-015, OGI CSE, 2003.
10. Y. Han. Automated Security Analysis of Internet Protocols using Coloured Petri Net Specification. Master of Science Thesis, Department of Electrical and Computer Engineering, Queen’s University, Ontario, Canada, 1996.
11. M. Jakobsson and A. Juels. Proofs of work and bread pudding protocols. In *the IFIP TC6 and TC11 Joint Working Conference on Communications and Multimedia Security (CMS 99)*, Sep 1999.
12. A. Juels and J. Brainard. Client Puzzles: A Cryptographic Defense Against Connection Depletion Attacks. In *the 1999 Network and Distributed System Security Symposium (NDSS '99)*, pages 151–165, San Diego, California, USA, Feb 1999.
13. W. Mao. Time-Lock Puzzle with Examinable Evidence of Unlocking Time. In *Proceedings of the 7th International Workshop on Security Protocols*, pages 95–102, London, UK, 2000. Springer-Verlag.
14. R. Moskowitz. The Host Identity Protocol (HIP). Internet Draft, Internet Engineering Task Force, Jun 2006. <http://www.ietf.org/internet-drafts/draft-ietf-hip-base-06.txt>.
15. R. L. Rivest, A. Shamir, and D. A. Wagner. Time-lock Puzzles and Timed-release Crypto. Technical Report TR-684, Massachusetts Institute of Technology, Cambridge, MA, USA, 10 Mar 1996.
16. W. A. Simpson. IKE/ISAKMP Considered Harmful. *USENIX*, 24(6), dec 1999.
17. J. Smith, J. M. González Nieto, and C. Boyd. Modelling Denial of Service Attacks on JFK with Meadows’s Cost-Based Framework. In *Fourth Australasian Information Security Workshop (AISW-NetSec'06)*, volume 54, pages 125–134, 2006.

Securing Data Accountability in Decentralized Systems

Ricardo Corin¹, David Galindo², and Jaap-Henk Hoepman²

¹ University of Twente, Enschede, The Netherlands
corin@cs.utwente.nl

² Institute for Computing and Information Sciences, Radboud University Nijmegen,
The Netherlands
{d.galindo, jhh}@cs.ru.nl

Abstract. We consider a decentralized setting in which agents exchange data along with usage policies. Agents may violate the intended usage policies, although later on auditing authorities may verify the agents' data accountability with respect to the intended policies. Using time-stamping and signature schemes, we design and analyze an efficient cryptographic protocol generating communication evidences, in such a way that an agent is *accountable* in our protocol only if the agent behaved honestly.

Keywords: applied cryptography, data accountability, timed communication evidence, decentralized systems.

1 Introduction

In many situations, there is a need to share data between potentially untrusted parties while ensuring the data is used according to given policies. For example, Alice may be interested in sending her e-mail address to Bob, but also attaching a non-disclosure policy, so that Bob may not disclose Alice's email to anyone else (see Figure 1(1)).

Of course, a priori nothing guarantees that Bob will actually follow the policy, making the enforcement of such policies a difficult problem. *Access and usage control* [JSSB97, SS94, PS02, BCFP03] are exemplary enforcement mechanisms. In these approaches, a trusted access control service arbitrates data access at the moment the request happens, something that can be sometimes overly restrictive and expensive.

Recently, a more flexible approach has been proposed [CEDH⁺04, CCD⁺05]. In this approach, after Bob receives Alice's e-mail address, Bob is free to violate the policy, for instance by sending Alice's e-mail to Charlie and including a "free-to-be-spammed" policy (see Figure 1(2)). However, it could happen that later on Bob is *audited* by an authority that requests a convincing proof of Bob's permission to disclose Alice's e-mail address. Auditing authorities are not fixed and pre-established; they may be formed dynamically by (groups of) agents that observe actions in the system. For example, consider that Alice starts getting tons of spam from Charlie (see Figure 1(3)). Alice may switch to "auditing

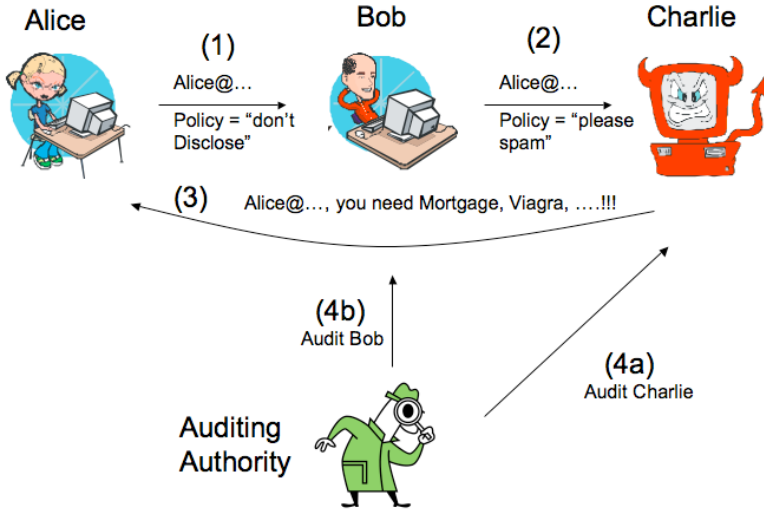


Fig. 1. An Illustrative Scenario

authority” mode, and enquiry Charlie for a proof of permission to use Alice’s e-mail address. A convincing proof exonerating Charlie would be evidence of the communication with Bob which happened before the spam, in which Charlie got Alice’s e-mail address along with a “free-to-be-spammed” policy. Upon seeing this, Alice can go to Bob and request a convincing proof he was allowed to disclose Alice’s e-mail to Charlie, which Bob can not provide and hence is found guilty of policy violation (see Figure 1(4a) and (4b)). In this case, we say that Bob violates the *datum accountability* of Alice’s e-mail address.

Following [CEdH⁺04, CCD⁺05], we allow the target system to be decentralized, and it may very well be the case that Alice, Bob and Charlie are registered in different domains, which are not even synchronized in their local clocks; even worse, each domain’s clock may not even be keeping real time, but may be simply a logical clock (e.g. a counter). Still, an auditing authority would need to recognize events (e.g. communications) that happened before others. For example, the fact that Charlie sent spam to Alice only after having received the permission from Bob exonerates Charlie. Also, an auditing authority needs to be able to recognize valid communication evidences from fake ones. Moreover, Bob must not be able to repudiate the communication with Charlie.

CONTRIBUTIONS. The works [CEdH⁺04, CCD⁺05] focus on designing a high-level formal proof system to build authorization permissions starting from the atomic communication evidences, which are abstracted away and assumed to be sound (this is analogous to Dolev-Yao models [DY83, CES06] which assume “perfect encryption” in the analysis of security protocols). However, from the above discussion it is clear that it is not trivial to design an effective and secure cryptographic protocol to achieve valid communication evidences to let auditing

authorities do their work, as we are in a decentralized setting in which different agents may collude in order to fool the auditing authorities.

In this paper, we design such a cryptographic protocol, which allows agents to exchange usage policies among them in such a way that valid communication evidence is generated after the exchange. Our protocol is efficient, and it's based on standard cryptographic building blocks, as detailed next.

BUILDING BLOCKS. We use two building blocks in our protocol: time-stamping and signature schemes. We choose to model different agents' domains running different local clocks, which may be logical, as described above. Our model is quite flexible, and in particular brings the following advantages:

- First, we allow the protocol to work in decentralized environments, in which clocks domains do not need to interact between them;
- Second, it is possible to use both digital-based and hash-based time-stamping schemes (see [ABSW01] for an overview), the choice depending on the use the implementor has in mind. For instance, if the implementor prefers to regard the time-stamping servers as non-trusted, auditable time-stamping schemes can be used without loss of security. Time-stamping is used to order the events within the domain, and enables our cryptographic protocol to safely logs the different domains' times of the sender and the receiver into the communication evidence. At a later stage, the auditing authority can use this information for the auditing procedure. Signatures by the agents are used to provide integrity, authentication and non-repudiability to the communication evidences.

Plan of the paper. We introduce preliminaries and define the security requirements of the protocol in Section 2. The building blocks of the protocol (i.e. time-stamping services and signature schemes) are introduced in Section 3. We present the protocol in Section 4, along with an (informal) security analysis in Section 5. Finally Section 6 concludes the paper.

2 The Setting

Potentially untrusted agents $a, b, c \in \mathcal{G}$ are sharing data objects $d \in \mathcal{D}$ which must be used according to policies $\phi \in \Phi$. The agents are not enforced to comply with the policies, but they can be audited at any time by an Auditing Authority AA, which will check if they were allowed to perform a certain action on a given piece of data d at a certain moment in time. How an agent c shows *accountability* for possessing or using a certain piece of data depends on the following. If the agent c is the *owner* of the data, he is allowed to perform any action on the data; otherwise, the agent must show that another agent b sent him a policy ϕ containing permissions to execute the actions under investigation. Eventually, the AA could iterate the process starting from agent b , and ideally the sequence should led to the owner of the piece of data.

We do not study how to build evidences for the creation of data. For the purpose of this paper, certifying that an agent owns/creates a piece of data needs the interaction with several organizations external to the system, since it involves legal and societal aspects (as an example, an organization issuing patents). Therefore, we assume there exists a public function

$$\text{OWNER} : \mathcal{D} \longrightarrow \mathcal{G}. \tag{1}$$

stating to whom the data belongs.

In our setting, $\text{comm}(t_1, t_2, a \Rightarrow b, \phi)$ provides a proof that a string ϕ was sent from a to b with respect to some local logical time values t_1, t_2 , where t_1 refers to the sender local time and t_2 refers to the receiver local time. The properties that such a proof should satisfy are as follows:

Meaningfulness. $\text{comm}(t_3, t_4, b \Rightarrow c, \phi')$ provides a proof that a string ϕ' has been communicated from b to c .

The reason for this property is trivial. We want to be sure that a communication evidence between Bob and Charlie can only be generated if Bob has communicated ϕ to Charlie.

Unforgeability. It is not feasible to create an evidence $\text{comm}(t_1, t_2, a \Rightarrow b, \phi)$ when the local logical clock for b is $t'_2 > t_2$.

Assume the auditing authority contacts Bob at Bob's logical time t'_2 and asks him for a permission for having sent Alice's e-mail address to Charlie. The unforgeability property implies Bob can not provide an exonerating evidence even in the unlikely case he colludes with Alice, since exoneration would require an evidence of the form $\text{comm}(t_1, t_2, a \Rightarrow b, \text{spread Alice's e-address})$ to be created at time t'_2 with $t_2 < t'_2$.

Liability. A valid $\text{comm}(t_3, t_4, b \Rightarrow c, \phi')$ implies b has committed itself at a logical time t_3 to send ϕ' to c .

That is, if Charlie shows an evidence $\text{comm}(t_3, t_4, b \Rightarrow c, \phi')$ showing that Bob sent him a permission to spam Alice at Bob's logical time t_3 , then Bob is liable for showing he had permission to communicate this policy before time t_3 .

Comparability. Any pair of communication evidences $\text{comm}(t_1, t_2, a \Rightarrow b, \phi)$ and $\text{comm}(t_3, t_4, a \Rightarrow b, \phi')$, with a, b, c agents, should be comparable with respect to b 's local time, i.e. communication evidences with origin or destination b are totally ordered with respect to b 's local logical time.

With this property, we ensure Bob can not show an evidence allowing him to execute a certain action, such that the order of action permission reception and action execution is undetermined.

3 Building Blocks

3.1 Time-Stamping Schemes

Time-stamping is an important data integrity protection mechanism the main objective of which is to prove that electronic records existed at a certain time.

Two major time-stamping protocols, the absolute (hash-and-sign) time stamps protocol and the linking protocol have been developed. In the former a time-stamping authority (TSA) signs the concatenation of the hashed message and the present time by using its private key. Therefore the TSA is completely trusted in this case.

In the linking protocol, a time-stamp token is the hash value of the concatenation of the present hashed message and the previous hash value. A verifier can check the validity of the token by using published values for the hash chain. In this case, the TSA is not necessarily trusted, and auditing techniques can be used to detect if the TSA eventually cheated. We stress this audit does not need to be performed by our auditing authority; it is sufficient the AA trusts the specific time-stamping auditors.

Time-stamping schemes have been used in business applications and are even included in international standards [ISO]. The major schemes in use are [sur, aut, dig].

The formal security notions for time-stamping schemes are still a subject under discussion. In the following, we quote the syntax and security properties of a time-stamping scheme from [BLSW05]. Notice that the definitions below are abstracted away, so that the different implementations discussed in the literature can be plugged into our protocol. Additionally, an audit functionality can be added to time-stamping schemes with the syntax above. This auditing functionality is used when the TSA is not unconditionally trusted, and it allows to check that the TSA is behaving as expected.

Definition 1. *A time-stamping scheme TS is capable of: (1) assigning a time-value $t \in \mathbb{N}$ to each request $x \in \{0, 1\}^k$, and (2) verifying whether x was time-stamped during the t -th (maybe logical) time unit). It consists of the following components:*

REPOSITORY – *a write only database that receives k -bits digests and adds them to a list \mathcal{D} . REPOSITORY also receives queries $t \in \mathbb{N}$ and returns $\mathcal{D}[t]$ if $t \leq |\mathcal{D}|$. Otherwise, REPOSITORY returns reject.*

STAMPER – *operates in discrete time variables called rounds. During a t -th round, STAMPER receives requests x and returns pairs (x, t) . Let L_t be the list of all requests received during the t -th round. In the end of the round, STAMPER creates a certificate $c = \text{Stamp}(x; L_t, L_{t-1}, \dots, L_1)$ for each request $x \in L_t$. Besides, STAMPER computes a digest $d_t = \text{Publish}(L_t, \dots, L_1)$ and sends d_t to the repository.*

VERIFIER – *a computing environment for verifying time stamps. It is assumed that VERIFIER has a tamper-proof access to REPOSITORY. On input (x, t) , VERIFIER obtains a certificate c from STAMPER, and a digest $d = \mathcal{D}[t]$ from REPOSITORY, and returns $\text{Verify}(x, c, d) \in \{\text{yes}, \text{no}\}$. Note that x can be verified only after the digest d_t is sent to REPOSITORY.*

CLIENT – *any application-environment that uses STAMPER and VERIFIER.*

A time-stamping scheme $\text{TS} = (\text{Stamp}, \text{Publish}, \text{Verify})$ must satisfy the following correctness property: $\text{Verify}(x, \text{Stamp}(x, \mathcal{L}), \text{Publish}(\mathcal{L})) = \text{yes}$ for every $\mathcal{L} = (L_t, \dots, L_1)$ and $x \in L_t$.

Security requirements

In the rest of the paper, an adversary is meant to be any probabilistic polynomial time algorithm. It is assumed that adversaries \mathcal{A} is able to corrupt STAMPER, some instances of CLIENT and VERIFIER. The REPOSITORY is assumed to be non-corrupting. After publishing d_t it should be impossible to add a new request x to L_t and prove to a VERIFIER that $x \in L_t$ by building up a certificate c . The following security conditions must be then required:

Definition 2 (Consistency). *A time-stamping scheme is consistent if for every PPT adversary \mathcal{A}*

$$\Pr[\mathcal{L} = (L_t, \dots, L_1, c, x) \leftarrow \mathcal{A}(1^k) \mid x \notin L_t, \text{Verify}(x, c, \text{Publish}(L_t, \dots, L_1)) = \text{yes}] \text{ is negligible.}$$

With the security notion below, we want that an adversary \mathcal{A} can not perform the following attack: \mathcal{A} publishes a value d which is not computed by using the Publish function and then, after obtaining a new randomly generated string x , finds a certificate that $\text{Verify}(x, c, d) = \text{yes}$.

Definition 3 (Security against random back-dating). *A time-stamping scheme is secure against random back-dating if for every polynomially unpredictable distribution \mathcal{D} on $\{0, 1\}^k$ and $(\mathcal{A}_1, \mathcal{A}_2)$ probabilistic polynomial time algorithms*

$$\Pr[(d, a) \leftarrow \mathcal{A}_1(1^k), x \leftarrow \mathcal{D}, c \leftarrow \mathcal{A}_2(x, a) \mid \text{Verify}(x, c, d) = \text{yes}] \text{ is negligible.}$$

3.2 Signature Scheme

Definition 1. *A signature scheme $\Sigma = (\text{KeyGen}, \text{Sign}, \text{VerSign})$ consists of three probabilistic polynomial time algorithms:*

- **KeyGen** takes as input a security parameter 1^k , and outputs a pair (vk, sk) , where sk is the secret key of the user, and vk is the matching verification key.
- **Sign** takes as input a message m and the secret key sk , and produces a signature σ .
- **VerSign** finally, the verification algorithm takes as input a message m , a signature σ and the verification key vk , and returns **true** if σ is a valid signature of m , and **false** otherwise.

A signature scheme enjoys the correctness property if it satisfies the following condition: if $\text{KeyGen}(1^k) = (sk, vk)$ and $\text{Sign}(m, sk) = \sigma$, then $\text{VerSign}(m, \sigma, vk) = \text{true}$. In this case, we say that (m, σ) is a valid message-signature pair.

The standard security notion for signature schemes was introduced in [GMR88] and it is called *existential unforgeability against chosen-message attacks*. A signature scheme $\Sigma = (\text{KeyGen}, \text{Sign}, \text{VerSign})$ is called secure in the latter sense if the success probability of any PPT adversary \mathcal{A} in the following game is negligible in the security parameter 1^k :

1. KeyGen(1^k) outputs (vk, sk) and the adversary is given 1^k and vk .
2. $\mathcal{A}(1^k, vk)$ has access to a signing oracle $\text{Sign}(sk, \cdot)$, which on input a message m outputs its signature $\sigma(m)$.
3. \mathcal{A} succeeds if it outputs a valid signature on a message not submitted to the signing oracle.

4 A Communication Evidence Protocol

Our goal is to design a decentralized protocol performing as less on-line operations as possible. Our system includes certification authorities CA trusted by the auditing authority, which will be used for keys authenticity and non-repudiation purposes. The AA can choose to trust/distrust the time-stamping servers. This election will determine which time-stamping schemes are accepted in the system (cf. the discussion in Section 3.1). The AA is an algorithm taking inputs from the agents, but it does not need itself to be in possession of any special input like a public key or similar.

Let us assume the existence of a set of time-stamping and certification authorities satisfying the trust requirements imposed by the AA, as well as the existence of a public board in which each user is inscribed in an unique TSA and CA. Let us denote by TSA_a , CA_a and TSA_b , CA_b the authorities in which a and b are respectively registered. A time-stamping scheme is used to provide temporal evidence to bit strings. In our setting consists each time stamping authority TSA_a runs a local time variable. We say that \mathcal{T}_a is the local time managed by TSA_a . Despite the local time variables are run in a decentralized manner, we are still able to define an irreflexive partial ordering (denoted by \prec) between events in the different \mathcal{T}_a 's. Notice that every local time variable has a total order $<$, i.e. the natural ordering in the set \mathbb{N} .

The partial order is defined in terms of the relation ' D_1 existed before D_2 ', where D_1, D_2 are strings. The relation is determined via time-stamps: a valid time-stamp certificate c^a issued on round t^a by TSA_a on a string m , implies m existed before round t^a was closed. As a consequence, a time-stamp c^b issued on round t^b by TSA_b on a string m' such that c^a is a substring of m' , implies that c^a existed before round t^b was closed. Summing up, the t^a -th round in TSA_a ended before the t^b -th round in TSA_b did so. This enables to establish the partial order $t^a \prec t^b$.

Any agent a in the system has a unique pair of matching verification/signature keys (vk_a, sk_a) corresponding to a secure signature scheme Σ_a , and it is registered in a *single* TSA. We will refer to this as a is registered in TSA_a . We assume that these keys are revocable. Therefore, agents have access to certification authorities CA which ensure the verification key vk_a belongs to a and provide revocation mechanisms.

A communication evidence $\text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi)$ must satisfy the requirements outlined in Section 2: meaningfulness, unforgeability, liability and comparability. The temporal tag includes two temporal values; a first value t_1^a will prevent non-repudiation by a (and therefore makes a liable for having permission

to communicate ϕ at ‘time’ t_1^a), and a second value t_2^b will refer to the moment in which b is allowed to use the policy ϕ . The value t_2^b will prevent forging evidences by b even if a and b collude together.

Definition 4 (Syntax of our communication evidence protocol). A communication evidence protocol is a pair of functions $\text{CE} = (\text{Create}, \text{Validate})$.

Create. The parties initiating a run of the protocol **Create** are two agents $a, b \in \mathcal{A}$, where a is willing to send a policy ϕ to b and b is willing to receive and therefore use this policy. The protocol additionally involves their respective time-stamping authorities $\text{TSA}_a, \text{TSA}_b$; and the certification authorities for a, b which are denoted by CA_a, CA_b . The output is a communication evidence $\text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi)$.

Validate – can be run by any agent in the system, and requires interaction with TSA_a and TSA_b and the certification authorities agents CA_a, CA_b . It takes as input a communication evidence $\text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi)$, and it returns **true** if comm is valid, and **false** otherwise.

4.1 Protocol Specification

In order to be able to prevent non-repudiation of communication evidences, we need to slightly modify the revocation mechanism used by the certification authority. In particular, the certification authorities must contact the TSA where the user is registered to time-stamp the revocation information for that user’s public key. In this way, it is possible to check if a communication evidence was created before either the sender’s verification key vk_a or the receiver’s verification key vk_b were revoked.

– **Create**(a, b, ϕ):

1. a signs the concatenated string (ϕ, b) using the scheme Σ_a and the signing key sk_a . Let σ_1 denote the signature thus obtained.
2. a sends σ_1 to **Stamper** $_a$, and gets back a valid stamp (t_1^a, c_1^a) when the t_1^a -th round is closed.
3. a sends $ev1 := (a, vk_a, b, \phi, \sigma_1, t_1^a, c_1^a)$ to b .
4. b verifies that:
 - (a) vk_a is a ’s verification key.
 - (b) vk_a was not revoked before TSA_a ’s local time t_1^a (this is done by interacting with CA_a and TSA_a).
 - (c) $\text{VerSign}((\phi, b), \sigma_1, vk_a) = \text{true}$.
 - (d) $\text{Verify}(\sigma_1, c_1^a, d_{t_1^a}) = \text{yes}$, where $d_{t_1^a}$ is the corresponding entry in REPOSITORY_a .

If everything is fine, then b proceeds to the next step. Otherwise, b does not use policy ϕ .

5. b signs $ev1$ using sk_b . Let σ_2 denote the signature thus obtained.
6. a sends σ_2 to **Stamper** $_b$, and gets back a valid stamp (t_2^b, c_2^b) when the t_2^b -th round is closed (and therefore $t_1^a < t_2^b$). Let $ev2 := (b, vk_b, \sigma_2, t_2^b, c_2^b)$.
7. Finally

$$\text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi) := (ev1, ev2).$$

- **Validate**($a, b, \phi, \text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi)$):
1. Contact TSA_b and get the value T_b of the current (non-closed) round.
 2. Verify that:
 - (a) $\text{VerSign}((\phi, b), \sigma_1, vk_a) = \text{true}$.
 - (b) vk_a is the a 's verification key.
 - (c) vk_a was not revoked before TSA_a 's local time t_1^a .
 - (d) $\text{Verify}(\sigma_1, c_1^a, d_{t_1^a}) = \text{yes}$, where $d_{t_1^a}$ is the corresponding entry in REPOSITORY_a .
 - (e) vk_b is the b 's verification key.
 - (f) vk_a was not revoked before TSA_b 's local time t_2^b .
 - (g) $\text{VerSign}(ev_2, \sigma_2, vk_b) = \text{true}$.
 - (h) $\text{Verify}(\sigma_2, c_2^b, d_{t_2^b}) = \text{yes}$, where $d_{t_2^b}$ is the corresponding entry in REPOSITORY_b .
 - (i) $t_2^b < T_b$.
- If every checking is correct, then return **true**. Otherwise, return **false**.

5 Security Analysis

Meaningfulness. “A string ϕ has been communicated from a to b ”

Firstly, if the communication evidence is verified in the positive, neither a 's or b 's verification keys were revoked beforehand. That a is the origin of the communication and b is the receiver, is guaranteed by two facts: on the one hand, a valid signature on the message (ϕ, b) can only be produced by a since we are using an unforgeable signature scheme Σ_a ; on the other hand, only b is able to compute the signature σ_2 for a similar reason, and he can do that only after a sends ev_1 to him.

Unforgeability. “It is not feasible to create an evidence $\text{comm}(t_1^a, t_2^b, a \Rightarrow b, \phi)$ when the local logical clock for b is set to t_2^b with $t_2^b > t_1^a$ ”

For creating such an evidence at time $t_2^b > t_1^a$, the adversary must break the security against random back-dating of the time-stamping scheme.

Liability. “ a commits itself to send ϕ to b at logical time t_1^a ”

a commits to message (ϕ, b) as soon as he signs it; a valid time-stamp (t_1^a, c_1^a) on σ_1 implies (ϕ, b) was signed before the local time counter at TSA_a was set to t_1^a . Therefore, a expresses at time t_1^a his willingness to transfer ϕ to b if he follows the protocol. Finally, if the communication evidence is verified in the positive, a 's verification key was not revoked beforehand.

Comparability. “communication evidences with origin or destination a are totally ordered with respect to a 's local logical time”

This is guaranteed by the fact that a 's logical time is \mathcal{T}_a and that \mathcal{T}_a has a total order by definition.

6 Conclusions

We define a cryptographic protocol to provide valid communication evidences, that can be used later by auditing authorities. Even though our protocol is

aimed towards guaranteeing data accountability in the settings of [CEdH⁺04, CCD⁺05], we believe that our protocol can be easily adapted to provide secure transport of arbitrary payloads in decentralized settings, where exchanges need to be logged with communication evidences recording the relative domain times in which the exchanges took place.

References

- [ABSW01] A. Ansper, A. Buldas, M. Saarepera, and J. Willemson. Improving the availability of time-stamping services. In *ACISP 2001*, volume 2119 of *Lecture Notes in Computer Science*, pages 360–375, 2001.
- [aut] <http://www.authentidate.com/>.
- [BCFP03] E. Bertino, B. Catania, E. Ferrari, and P. Perlasca. A logical framework for reasoning about access control models. *ACM Transactions on Information and System Security (TISSEC)*, pages 71–127, 2003.
- [BLSW05] A. Buldas, Peeter Laud, M. Saarepera, and J. Willemson. Universally composable time-stamping schemes with audit. In *ISC 2005*, volume 3650 of *Lecture Notes in Computer Science*, pages 359–373, 2005.
- [CCD⁺05] J. G. Cederquist, R. J. Corin, M. A. C. Dekker, S. Etalle, and J. I. den Hartog. An audit logic for accountability. In *6th Int. Workshop on Policies for Distributed Systems & Networks (POLICY)*, pages 34–43. IEEE Computer Society Press, 2005.
- [CEdH⁺04] R. Corin, S. Etalle, J. I. den Hartog, G. Lenzini, and I. Staicu. A logic for auditing accountability in decentralized systems. In *Proc. of the 2nd IFIP Workshop on Formal Aspects in Security and Trust (FAST)*, volume 173, pages 187–202. Springer, 2004.
- [CES06] R. Corin, S. Etalle, and A. Saptawijaya. A logic for constraint-based security protocol analysis. In *IEEE Symposium on Security and Privacy*, 2006.
- [dig] <http://www.digistamp.com/>.
- [DY83] D. Dolev and A.C. Yao. On the security of public key protocols. *IEEE Transactions on Information Theory*, 29(2):198–208, 1983.
- [GMR88] S. Goldwasser, S. Micali, and R.L. Rivest. A digital signature scheme secure against adaptive chosen-message attacks. *SIAM J. Comput.*, 17(2):281–308, 1988.
- [ISO] ISO IEC 18014-3,time-stamping services part 3: Mechanisms producing linked tokens.
- [JSSB97] S. Jajodia, P. Samarati, V. S. Subrahmanian, and E. Bertino. A unified framework for enforcing multiple access control policies. In J. Peckham, editor, *SIGMOD 1997, Proc. International Conference on Management of Data*, pages 474–485. ACM Press, 1997.
- [PS02] J. Park and R. Sandhu. Towards usage control models: Beyond traditional access control. In E. Bertino, editor, *Proc. of the 7th ACM Symposium on Access Control Models and Technologies (SACMAT)*, pages 57–64. ACM Press, 2002.
- [SS94] R. Sandhu and P. Samarati. Access control: Principles and practice. *IEEE Communications Magazine*, 32(9):40–48, 1994.
- [sur] <http://www.surety.com/>.

Privacy Friendly Information Disclosure

Steven Gevers and Bart De Decker

Department of Computer Science, K.U.Leuven,
Celestijnenlaan 200A, B-3001 Leuven, Belgium

Abstract. When using electronic services, people are often asked to provide personal information. This raises many privacy issues. To gain the trust of the user, service providers can use privacy policy languages such as P3P to declare the purpose and usage of this personal information. User agents can compare these policies to privacy preferences of a user and warn the user if his privacy is threatened. This paper extends two languages: P3P and APPEL. It makes it possible to refer to certified data and credentials. This allows service providers to define the minimal level of assurance. It is also shown how different ways of disclosure (exact, blurred, verifiably encrypted, ...) can be specified to achieve more privacy friendly policies. Last, the paper describes a privacy agent that makes use of the policies to automate privacy friendly information disclosure.

1 Introduction

When using electronic services, people are often asked to provide personal information. This raises many privacy issues. To gain the trust of the user, service providers can use privacy policy languages to specify the purpose and usage of this personal information. User agents can compare these policies to privacy preferences of a user and warn the user if his privacy is threatened. Two well known privacy languages are P3P (The Platform for Privacy Preferences [14]) and APPEL (A P3P Preference Exchange Language [15]). The former is used for privacy policies, the latter for privacy preferences.

There are three ways in which a user can prove personal information to a service provider. He can provide it as uncertified data, certified data or embedded in a credential. In this paper, the term *information structure* denotes all three. *Uncertified data* is data that is not certified by another entity. It can easily have been stolen, forged or made up. *Certified data* is data that is endorsed (e.g. signed) by a certifying entity. When a service provider receives certified data, he can be sure that the information is correct. However, when someone receives certified data, he can easily pretend to be the legitimate owner. *Credentials* contain certified data and offer a means to ensure the service provider that the person sending the data is indeed the one to whom the credential was issued. Examples are X.509 certificates [13] and private credentials [1, 2]. Hence, credentials offer the highest assurance and allow for implementing secure services.

Private credentials have many privacy friendly properties. They make it possible to hide the values that the service provider does not need to know (selective disclosure). Also, they support different *ways of disclosure*. It is possible to prove properties of attributes (e.g. proving being over eighteen) or just proof possession of the credential without revealing it.

The paper extends P3P and APPEL with the different ways of disclosure and the different types of information structures. This way, service providers can request a certain level of assurance. Also, the user's privacy can be better protected. The policies make it possible to automate personal information disclosure. The paper describes how they are used by a *privacy agent*. The privacy agent discloses information structures to a service provider. User intervention is only required when absolutely necessary. However, the user's privacy is protected according to his preferences. More detailed information about our approach can be found in [4].

Next section extends P3P and APPEL with information structures and the different ways of disclosure. Section 3 gives a description of the privacy agent. Section 4 discusses our approach. Last section gives some conclusions.

2 Extending Privacy Languages

This section extends P3P and APPEL. The extensions help service providers to define which (parts of) information structures they are willing to accept. Users can define accurate privacy preferences about their information structures. This section assumes basic knowledge of P3P and APPEL. A short introduction can be found in [4].

2.1 Information Structure Description Language

This section introduces a language that is able to define information structures in XML. By including these descriptions in privacy policies and privacy preferences it is possible to state which information structures may be used. The language makes it possible to reason about the contents of different types of information structures in a uniform way. Uncertified data keeps information about the owner of the data. Certified data and credentials also include information about the certifying entity. Furthermore, they may have several properties. An information structure description can thus be divided in one part about the owner, another part about the certifier and yet another part about properties. Figure 1 shows the description of a private credential containing the age and name of its owner. More examples can be found in [4].

The *name*-attribute of the <INFORMATIONSTRUCTURE>-tag is used to allow for making references to it. *Type* defines the type of information structure. Currently, *privatecredential*, *X.509Certificate*, *uncertifieddata* and *certifieddata* are defined. A generic value *credential* can be used to denote both private credentials and X.509 certificates.

The <OWNER>-part in figure 1 defines properties (attributes) of the owner. In this case, the credential includes information about the person's age and name.

```
<INFORMATIONSTRUCTURE name="IDcredential" type="credential">
  <OWNER>
    <PROPERTY name="name">Bill</PROPERTY>
    <PROPERTY name="age">18</PROPERTY>
  </OWNER>
  <CERTIFIERCREDENTIAL name="municipalityCA"/>
  <PROPERTIES>
    <PROPERTY name="notValidAfter">03-06-07 01:01:01</PROPERTY>
  </PROPERTIES>
</INFORMATIONSTRUCTURE>
```

Fig. 1. The information structure description language

Credentials are verified through the use of other credentials. This leads to a chain of credentials that ends with a root credential (e.g. X.509 certificate chains). The <CERTIFIERCREDENTIAL>-tag points to an information structure description of the next credential in the chain. The <ROOTCREDENTIAL>-tag defines the root credential.

The last part defines the properties of the information structure. Certified data typically has a signature algorithm as property. Common properties of credentials are, for example, the validity period and revocation information.

It is important that every information structure is mapped on the tags in a consistent way. That way it is possible to use the different types of information structures interchangeably.

2.2 Privacy Policies

Whenever a service provider needs personal information from the user, he has to create a privacy policy. To include information structures in P3P the service provider first has to define the information structures he is willing to accept from users. He can do this by using the information structure description language, described in 2.1. Then, the service provider is able to make references to these descriptions in the P3P policy. To make references, two parts of P3P policies are extended: the <DATA-DEF> and the <DATA-GROUP> part. The former is used to define data elements. The latter is used to specify the information that has to be disclosed. Note that P3P privacy policies also contain parts about, for example, purpose and retention time (e.g. the address of a user is required for sending advertisements). Our approach does not change these parts. Unchanged parts that are irrelevant for the examples are not included in this paper.

```

<INFORMATIONSTRUCTURE name="IDcredential" type="credential">
  <OWNER>
    <PROPERTY name="name"/>
    <PROPERTY name="age"/>
  </OWNER>
  <CERTIFIERCREDENTIAL name="municipalityCA"/>
</INFORMATIONSTRUCTURE>

```

Fig. 2. Information structure descriptions

Defining Acceptable Information Structures. The service provider has to create information structure descriptions containing the distinguishing tags of the information structure he is willing to accept. An information structure matches a description if it contains *at least every tag* in the description. The user is allowed to use every information structure that matches the description.

Figure 2 shows an information structure description of a credential. If the service provider requests an *IDcredential*, every credential that contains the attributes *age* and *name* can be used. Furthermore, the credential has to be certified by the municipality CA. The `<CERTIFIERCREDENTIAL>`-tags points to an information structure description of a credential of this entity. In [4], a *level*-attribute is described that makes it possible to refer to credentials higher in the credential chain.

Note that it is possible to point to multiple information structures with only one description. Private credentials as well as X.509 certificates can match the specification in figure 2. This is an important aspect in this paper. Different types of information structures can be used interchangeably. Services can be made more accessible by allowing users to show their personal information in different ways.

Creating References to Information Structures. In P3P, the `<DATA-DEF>`-tag is used to define data elements. This tag is extended to be able to refer to attributes of information structures. Figure 3 shows that if the service provider requests a statement on the *age* of a user, only the age attribute of

```

<DATA-DEF name="age" short-description="The age of the client">
  <INFORMATIONSTRUCTURE name="IDcredential">
    <OWNER>
      <@PROPERTY name="age"/>
    </OWNER>
  </INFORMATIONSTRUCTURE>
</DATA-DEF>

```

Fig. 3. References to information structures

the *IDcredential* has to be used. The '@' indicates the parts of the information structure a user has to disclose. Hence, when using private credentials, the owner may hide the other attributes.

Defining Different Ways of Disclosure. In standard P3P a service provider can only state he needs certain information in clear text. With private credentials, however, there are more possibilities.

- It is possible to show/prove the value of credential attributes (i.e. *clear text*).
- Users are able to *prove knowledge* of a certain attribute. This corresponds to proving ownership of a credential containing the attribute without revealing any attributes.
- A user can disclose information as a *verifiable encryption*. A verifiable encryption is associated with a condition and a third party. If the condition is fulfilled, the third party is allowed to decrypt the encrypted information. This can, for example, be useful to identify a person in case of abuse. Cryptographic mechanisms ensure that the encryption contains the information requested by the service provider.
- A user can prove *equations* (\leq , \geq , $<$, $>$, \neq and $=$). It is possible to compare an attribute with a *known value* or with *attributes of other credentials*. These equations can be relatively complex: e.g. $attr1 + 7 \leq attr2 \cdot (4 + attr3)$. Service providers frequently request the interval a certain attribute belongs to. An example can be a site that needs to know that the user's income is in a certain interval. These intervals are typically defined by a startpoint and a step. The user then has to provide a number k and prove $start + k \cdot step \leq attribute \leq start + (k + 1) \cdot step$.

In order to allow the service provider to define how personal information must be shown, the <DATA-GROUP> part of P3P policies is extended. Figure 4 defines that the user can either use his *IDcredential* and prove being over eighteen or proof knowledge of a VISA card credential (i.e. proof to be the owner of a valid VISA card credential). The information will be used for pseudonymous analysis. The VISA card credential is not worked out in this paper.

Additional tags are introduced to support the different ways of disclosure. The tags <AND> and <OR> can be used to define more complex policies such as the ones in [3]. More information and examples about the different tags can be found in [4]. Our approach allows for a distinction between the personal information requested by the service provider and the technologies that are used to realize the disclosure. For instance, if a user's *IDcredential* is a private credential, a zero knowledge proof can be used to prove his adulthood. If he owns an X.509 certificate, he has to show the entire certificate and disclose *every* attribute in clear text.

The *level of disclosure* is a partial relation based on the different ways of disclosure and the level of assurance provided by the types of information structures. x has a higher level of disclosure than y if x reveals, in *every* case, more

```

<POLICIES>
  <POLICY name="analysisPolicy">
    <STATEMENT>
      <PURPOSE><pseudo-analysis/></PURPOSE>
      <DATA-GROUP>
        <OR>
          <GT>
            <DATA ref="#age"/>
            <VALUE>18</VALUE>
          </GT>
          <PROOFOFKNOWLEDGE>
            <DATA ref="#VISAcard">
          </PROOFOFKNOWLEDGE>
        </OR>
      </DATA-GROUP>
    </STATEMENT>
  </POLICY>
</POLICIES>

```

Fig. 4. Extending a data group in P3P

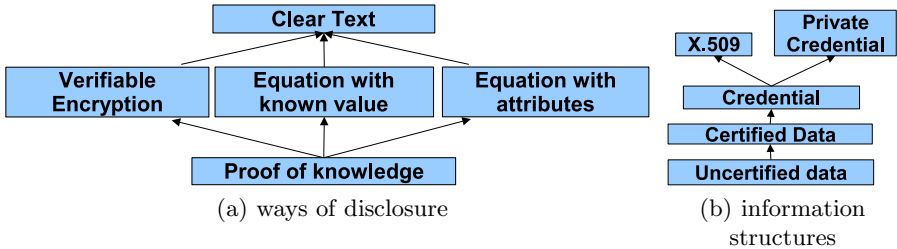


Fig. 5. Levels of disclosure

information than y . For example, showing an attribute in clear text reveals more information than proving an equation. Also, the type of information structure used must provide at least an equal level of assurance. x has a higher (or equal) level of disclosure than y if it has at least the same level in both figure 5(a) and 5(b). The level in which the service provider wants to receive the attributes is the *minimum level of disclosure*. When a user has to disclose a property of an attribute, the service provider will of course accept it if the user gives it away in clear text. Moreover, he will also accept it if the user discloses more information than necessary (for example, by using X.509 certificates). Also, if the service provider requests uncertified data, he will also accept a credential (assuming he is capable of handling the protocols associated with that type of information structure).

2.3 Privacy Preferences

The service provider has specified which information he wants to receive and how the user should provide it. APPEL can be similarly extended as P3P to include information structures. First, the user needs to have descriptions of his information structures. If an information structure has well defined semantics (e.g. X.509 certificates), it is possible to generate the tags of its description automatically [4]. If its semantics are not specified, it is impossible to automate the generation of information structure descriptions. In this case, the entity that issued the information structure could provide the description. References to these descriptions are made similarly as in extended P3P. Figure 6 specifies that it is allowed (behavior='request') to disclose being older than 18 for pseudonymous analysis.

```

<appel:RULE behavior="request">
  <p3p:POLICY>
    <p3p:STATEMENT>
      <p3p:PURPOSE><p3p:pseudo-analysis/></p3p:PURPOSE>
      <p3p:DATA-GROUP>
        <p3p:DATA ref="#age">
          <p3p:GT>
            <p3p:VALUE>18</p3p:VALUE>
          </p3p:GT>
        </p3p:DATA>
      </p3p:DATA-GROUP>
    </p3p:STATEMENT>
  </p3p:POLICY>
</appel:RULE>

```

Fig. 6. Defining how an attribute can be disclosed

The level at which the user wants to show attributes is the *maximum level of disclosure*. In the example, the user will not reveal his age in clear text. Comparisons can be made weaker. The user will allow to prove being older than sixteen because this reveals less information than if he proves being over eighteen. However, he will not prove being older than, for example, thirty.

3 A Privacy Agent

The previous section explained how P3P and APPEL can be extended with different types of information structures and different ways of disclosure. This section describes a *privacy agent* that shows the benefits of these extensions. When a user wants to use a service, a privacy policy is sent to the privacy agent. The privacy agent first checks whether the user's privacy preferences allow the

disclosure. Then, the privacy agent searches the user's information structures to find the ones that can be used to fulfil the privacy policy. After that, the privacy agent can either start the information disclosure automatically or show the (combinations of) information structures that can be used to the user. The latter is comparable to the identity selector of Microsoft CardSpace [8]. The user can then choose how his personal information has to be disclosed.

For the privacy agent to be successful, it has to be *user friendly*. One aspect is privacy preferences. Most users are not able to generate complex privacy preferences. To handle this, the privacy agent retrieves privacy preferences from a trusted third party. This approach is described in [5]. Also, if the user has a choice between several (combinations of) information structures, the privacy agent tries to help the user. By using *sensitivities* (described in [4]) the most privacy friendly combination is calculated. This combination is suggested to the user to help him protect his privacy.

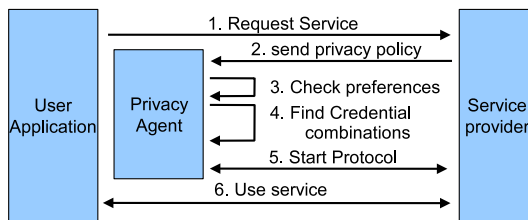


Fig. 7. Description of the system

Figure 7 shows the interaction. The following steps are necessary:

1. The user application requests the service.
2. The service provider sends his privacy policy to the privacy agent.
3. The privacy agent checks whether the policy matches the privacy preferences of the user.
4. The privacy agent selects the necessary credentials that comply with the policy. This action comprises more than just selecting information structures that contain the requested attributes. For example, when using X.509 certificates, all included attributes are revealed which may be more than necessary. The privacy preferences must be checked again to see whether it is allowed to show every attribute included. If there are different possibilities to show the information, the most privacy friendly combination should be calculated (using *sensitivities* [4]). Based on the user's preferences, the privacy agent can either start the information disclosure automatically or request the user's consent.
5. A protocol is started to disclose the information. First, the privacy agent informs the service how the information will be shown. When this information is exchanged, the correct protocol can be started. This approach makes it

necessary for the service provider to support the appropriate protocols for the different information structures. To deal with this, a protocol compiler as mentioned in [3] can be used.

6. The user application can make use of the service.

4 Discussion

The extensions of the language provide several advantages. Including the different ways of disclosure allows for more privacy friendly policies. It makes it possible to include the properties of private credentials. Other privacy languages, such as Rei [10], EPAL [12] and XPref [9], are not able to do this. Using the different types of information structures, service providers can define a certain level of assurance in their policy. It also makes it possible to automatically select the information structures that can be used for the personal information disclosure.

[3] proposes a language that allows one to specify what data to release and how to release it. Their specifications can be converted to the XML notation proposed in this paper. Our work extends the functionality by allowing the user to use different types of information structures. Our work also uses privacy preferences to check whether the personal information disclosure is allowed.

Usability is an important concern for the privacy agent. Almost every aspect of personal information disclosure can be automated. This way, complex technologies (e.g. private credentials) can be used without even understanding their basics.

The privacy agent is very *extendable*. The information structure description language is very general. This makes it possible to include all types of information structures in the system by mapping them on the language.

A positive aspect of our approach is that it can be very useful for the service provider. A service provider can easily give the user many options to prove personal information. For example, in figure 4, if a user does not want to prove knowledge of a VISA card, he can still prove his age with his IDcredential. This makes services more *accessible*. By making use of credentials, services can be made more secure.

The impact of the extensions on the evaluator of the privacy preferences is rather limited. Policies that give the users a choice are split into multiple policies. Checking the levels of disclosure does not require complex calculations either.

Microsoft CardSpace [8] provides a mechanism similar to our work. However, it does not include privacy policies. It also does not include the different ways of disclosure. Only clear text claims can be handled. Our approach is able to put constraints on the properties of information structures that are allowed. This can, for example, be useful if a site wants his users to have a passport that will remain valid for at least six months. Note that the Microsoft CardSpace fits perfectly in our system. Infocards can be described using the information structure description language. Claims are properties of the owner; the reference to the security token service is a property of the information structure itself. Requested claims can easily be included in a predefined privacy policy. Our approach is also

more general than Microsoft CardSpace. CardSpace always needs to contact a security token service to obtain a credential. Our approach is able to use credentials that are fully under the control of the user such as, for example, credentials on an electronic identity card.

Our privacy agent provides more functionality than existing user agents focussing on privacy policies such as AT&T's privacy bird [6] and JRC P3P Proxy [7]. Both only warn users when a site does not respect their preferences. Our privacy agent is able to use the different information structures and to define different ways of disclosure which makes it much more useful.

To make the system more usable, the system can be extended to support *trust negotiation*. Instead of sending a privacy policy to the privacy agent, the trust-target graph procedure discussed in [11] can be used. The extensions of the privacy languages proposed in this paper can be used to support the communication between the different parties. This is future work.

5 Conclusions

This paper proposed two extensions to P3P and APPEL. It is possible to include different types of information structures and different ways of disclosure. A privacy agent is described that makes use of these policies. The privacy agent provides user and privacy friendly information disclosure.

References

1. J. Camenisch and E. Van Herreweghen: Design and Implementation of the Idemix Anonymous Credential System. In *Proc. 9th ACM Conf. Computer and Comm. Security*, 2002.
2. S. Brands: Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy, 2000.
3. J. Camenisch, D. Sommer and R. Zimmermann: A general certification framework with applications to privacy-enhancing certificate infrastructures. Tech. Rep. RZ 3629, IBM Zurich Research Laboratory, July 2005.
4. S. Gevers and B. De Decker: Automating privacy friendly information disclosure. Tech. Rep. CW441, Katholieke Universiteit Leuven, May 2006
5. G. Yee and L. Korba: Semi-Automated Derivation of Personal Privacy Policies. In *IRMA '04: Proceedings of the 2004 Information Resources Management Association International Conference*, 2004.
6. AT&T Privacy Bird. <http://www.privacybird.com/>
7. JRC P3P Resource Centre. <http://p3p.jrc.it/>
8. Microsoft CardSpace <http://msdn.microsoft.com/winfx/reference/infocard/default.aspx>
9. R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu: An XPath based preference language for P3P. In *Proc. of the 12th Int'l World Wide Web Conference*, 2003.
10. L. Kagal, T. Finin, and A. Joshi: A policy based approach to security for the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, 2003.

11. J. Li, N. Li and W.H. Winsborough: Automated trust negotiation using cryptographic credentials. In *Proceedings of the 12th ACM Conference on Computer and Communications Security*, 2005.
12. The Enterprise Privacy Authorization Language (EPAL 1.1) <http://www.zurich.ibm.com/security/enterprise-privacy/epal/>
13. R. Housley, W. Ford, W. Polk and D. Solo: RFC 2459: Internet X.509 Public Key Infrastructure Certificate and CRL Profile
14. Platform for Privacy Preferences (P3P) Project. <http://www.w3.org/P3P/>
15. A P3P Preference Exchange Language 1.0 (APPEL1.0). <http://www.w3.org/TR/P3P-preferences/>

KSinBIT 2006 PC Co-chairs' Message

The impact of the upcoming Internet on scientific research worldwide has been enormous, not least in biomedical research. The Human Genome Project in particular was the inspiration for many biological databases publicly available via the Internet. As of now, conducting biomedical research without the Internet is nearly impossible. The information needed for analysis and interpretation of experimental results is usually scattered over a multitude of heterogeneous data sources: sequence databases, protein resources, gene expression data repositories, literature databases, functional annotation databases, etc. Many researchers depend on the Internet as the most important source of biomedical information. As the amount of available data increases at a rate never seen before, researchers are now faced with the problem of finding the information they need, in a format they can work with.

Several initiatives exist that try to integrate multiple data sources or facilitate complex bioinformatics queries and analyses. However, the integration is not always in tune with the user. The aim of this workshop was to bring together researchers and practitioners to exchange ideas with respect to knowledge systems in bioinformatics that make extensive use of medical and biological semantics and ontologies, Web services technologies, or distributed databasing and computing to tackle the issues mentioned above. Out of approximately 20 submitted papers, 10 papers were accepted for oral presentation. These papers are published in this volume. Together they give a very nice overview of where we are now and where current research is headed. We were very pleased with the number of people interested in contributing to this workshop, both authors and reviewers. Every submission received three outstanding reviews, which made the task of accepting papers of high quality quite easy. We hope you will enjoy reading them.

August 2006

Maja Hadzic, Curtin University of Technology, Australia
Bart De Moor, Katholieke Universiteit Leuven, Belgium
Yves Moreau, Katholieke Universiteit Leuven, Belgium
Arek Kasprzyk, European Bioinformatics Institute, UK

Suggested Ontology for Pharmacogenomics (SO-Pharm): Modular Construction and Preliminary Testing

Adrien Coulet^{1,2}, Malika Smaïl-Tabbone²,
Amedeo Napoli², and Marie-Dominique Devignes²

¹ KIKA Medical,

35 rue de Rambouillet, 75012 Paris, France

² LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP),
Campus scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France

Abstract. Pharmacogenomics studies the involvement of interindividual variations of DNA sequence in different drug responses (especially adverse drug reactions). Knowledge Discovery in Databases (KDD) process is a means for discovering new pharmacogenomic knowledge in biological databases. However data complexity makes it necessary to guide the KDD process by representation of domain knowledge. Three domains at least are in concern: genotype, drug and phenotype. The approach described here aims at reusing whenever possible existing domain knowledge in order to build a modular formal representation of domain knowledge in pharmacogenomics. The resulting ontology is called SO-Pharm for Suggested Ontology for Pharmacogenomics. Various situations encountered during the construction process are analyzed and discussed. A preliminary validation is provided by representing with SO-Pharm concepts some well-known examples of pharmacogenomic knowledge.

1 Introduction

Pharmacogenomics is the study of genetic determinants of drug responses. It involves relationships between at least three actors of interindividual differences in drug responses: genotype, drug, and phenotype (Fig. 1)[1]. Relevant genotype features are mostly genomic variations and particularly Single Nucleotide Polymorphisms (SNP). The latter are one-nucleotide substitutions occurring in a studied population with a minimum frequency of 1 %. Such genomic variations modulate drug effect, and have consequences on individual phenotype from the microscopic level (gene expression, protein activity, molecule transport, etc.) to the macroscopic level (clinical outcomes, etc.).

At present, best-recognized and completely developed examples of genomic variations altering drug response in human are monogenic traits acting on drug metabolism. Nevertheless, description of complex polygenic systems has recently proven that regulatory networks and many non genetic factors (e.g., environment, life style) also influence the effect of medications. Consequently, the discovery of new pharmacogenomic knowledge is a challenging task that necessitates

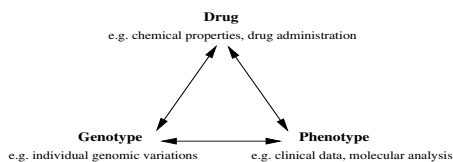


Fig. 1. Triangular schematization of the pharmacogenomic domain

the management of complex data. For example, the design of a clinical trial in pharmacogenomics relies on the selection of genes involved in drug response, selection of associated relevant genomic variations and on knowledge about the phenotypes associated with these genomic variations [2]. An interesting research direction is the integration of biological data stored in public annotated biological data banks, and clinical data resulting from clinical trials. This integration may allow, in a second stage, the discovery of pharmacogenomic knowledge thanks to the KDD process.

The KDD process is aimed at extracting from large databases information units that can be interpreted as reusable knowledge units (such as RDF/OWL triples). This process is based on three major steps: (a) the datasets are extracted from selected data sources and prepared for data mining, (b) are mined (with symbolic or numerical methods), finally, (c) the extracted information units are interpreted by domain expert to become reusable knowledge units [3]. All along this process, domain knowledge, embedded within an ontology, can be used to guide the various steps:

- a) During the preparation step it facilitates integration of heterogeneous data.
- b) During the mining step, domain knowledge guides the filtering of input and output data.
- c) In the interpretation step, it helps the experts for reasoning on the extracted units.

In order to achieve KDD in pharmacogenomics, we decided to develop a knowledge-based approach and therefore to explicit domain knowledge within an ontology. More and more biomedical ontologies are being developed today and often cover overlapping fields. To favor reuse of and access to ontologies, most biological ontologies are freely available. For instance, the Protégé ontologies library [4] provides various formal ontologies and the Open Biomedical Ontologies (OBO) portal [5] gathers many controlled vocabularies for the biomedical domain. Although the associated ontology is not available, the PharmGKB project has led to the construction of a valuable structured repository for pharmacogenomic data, aimed at catalyzing scientific research in this domain [6]. It provides a data model and a partial vocabulary for genotype and phenotype data of individuals involved in pharmacogenomic studies. In previous work, we developed the SNP-Ontology as a formal representation of genomic variation domain [7].

This paper describes the construction process of a “Suggested Ontology for pharmacogenomics” (SO-Pharm), that reuses existing ontologies designed for

pharmacogenomics sub-domains: genotype, drug, and phenotype. Section 2 describes the method used to build SO-Pharm and its content. Section 3 presents a preliminary testing of the ontology thanks to assertions of some established pharmacogenomic knowledge. Section 4 concludes on the work.

2 SO-Pharm Construction

2.1 Methodology Choice

Semiautomatic methods such as classification, itemset search, association rule extraction, text mining can be employed for ontology construction [8]. However, a manual construction is preferred here because of the objectives assigned to the ontology. In addition, the complexity of the field has favored a close collaboration with domain experts, nicely compatible with manual construction. Indeed, one difficulty consists in choosing and defining adequate concepts and properties for expressing pharmacogenomic knowledge. Manual construction is associated here with the use of a clearly defined methodology. Outlines of iterative processes for ontology construction have been described in [9,10,11]. We adapt here these methodologies to the case of pharmacogenomics, based on four steps:

- (i) specification, embedding definition of ontology domain and scope;
- (ii) conceptualization, that includes definition of list of terms and of concepts, and their articulation with existing ontologies;
- (iii) formalization, i.e., the translation of the conceptualization in a knowledge representation formalism (e.g. description logics);
- (iv) implementation, i.e., coding the formalized ontology in a knowledge representation language (e.g. OWL).

In the next sections we analyze and discuss the original orientations adopted during the SO-Pharm construction process.

2.2 Construction Issues

Specification. Domain and scope of SO-Pharm are primarily defined as follows. The domain considered should cover pharmacogenomic clinical trials. The ontology has to precisely represent individuals and groups of individuals involved in trials, their genotype, their treatment, their observed phenotype and the potential pharmacogenomic relations discovered between these concepts. Currently, SO-Pharm concepts do not cover epigenotype features, regulatory networks or metabolic pathways. SO-Pharm scope is to guide KDD in pharmacogenomics. According to the various steps of KDD process, SO-Pharm should reveal helpful in the following situations:

- integrating complementary data from various scopes: e.g. protein annotations and enzyme activity measurement;
- reconciling heterogeneous data: e.g. heterogeneous descriptions of genomic variations pertaining from locus specific databases and dbSNP;

- guiding data mining: for instance selection of a given class of genomic variations according to relations between these variations and the focus of the study;
- expressing data mining results as knowledge units: in order to compare with existing knowledge units and to infer new knowledge units;
- reusing of discovered pharmacogenomic knowledge: e.g. knowledge sharing between several independent projects.

During the specification step some strict nomenclature guidelines (e.g., for naming classes, associations, concepts, properties) are defined for the whole construction process. Then lists of domain terms are established. In the case of SO-Pharm ontology, the domain expert constitutes four primary term lists thanks to his own knowledge regarding respectively clinical trial, genotype, treatment, and phenotype descriptions. In parallel, data or knowledge resources in the domain are listed. These highly heterogeneous resources, including conceptual data model (in UML or UML-like), XML schemas, databases, ontologies, controlled vocabularies are displayed in Table 1 (*n.b.*: * are OBO ontologies). The study of their structure and content allows to considerably enrich the term lists.

The previous resource list is then refined for selecting relevant reusable knowledge resources according to following criteria (Table 2). First, it has been decided to take into account OBO ontologies, which are mostly used and known. Second, we have preferred the ontologies involved in the OBO-Foundry project that tries to adopt quality principles in ontology development [12]. The current resource list may be extended in the future and enriched with other interesting resources such as GO, Pathway Ontology, NCI, eVOC, Amino Acid Ontology, GandrKB.

Conceptualization. A UML class diagram is used here for representing the conceptual model of SO-Pharm. Term lists are exploited to identify ontology concepts which are assigned a name and a precise definition (free text). In SO-Pharm, a clinical item (or clinical data, or item) is defined as the measurement of a quantity for a given person, during a particular event, according to a measurement method. As well, a drug is composed of chemical compounds and may be included in a drug treatment and may have a commercial name. When concepts are identified, their hierarchical and non-hierarchical (i.e. object properties) relations are modeled by UML class diagrams. These diagrams are well adapted for conceptualization of domain knowledge because of their expressiveness and openness [13]. Fig. 2, 3 and 4 display UML class diagrams designed during SO-Pharm construction.

Articulation between the SO-Pharm concepts and external ontologies concepts is also established during this step (see Table 2 for prefix legend in UML class diagrams). The kind of relation (i.e. *embedding* or *extension*) invoked for reusing an ontology depends on its type [10]. Indeed, the majority of ontologies in biomedical domain may be organized into three categories: *meta-ontologies* providing domain-independent concepts and properties to be used as compounds for more specific ontologies (e.g. DOLCE, SUMO); *domain reference ontologies* representing a particular domain of reality and sorting entities of the domain

Table 1. List of explored resources for constructing term lists of the various domains

<i>Resource name</i>	<i>Resource type</i>	<i>Domain</i>	<i>URL</i>
dbSNP	XML schema, data model	genotype	http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap	XML schema	genotype	http://www.hapmap.org/
HGVBase	DTD, data model	genotype	http://hgvdbase.cgb.ki.se/
OMIM	Data resource	genotype, phenotype	http://www.ncbi.nlm.nih.gov/omim/
OMG SNP	Data model	genotype	http://www.omg.org/technology/documents/formal/snp.htm
MECV	Controlled vocabulary	genotype	http://www.ebi.ac.uk/mutations/
PharmGKB	XML schema, data model	genotype, drug, phenotype	http://www.pharmgkb.org/
Pharmacogenetics Ontology	Controlled vocabulary	genotype, phenotype	http://www.pharmgkb.org/home/projects/project-po.jsp
Sequence Ontology	Controlled vocabulary*	genotype	http://song.sourceforge.net/
Gene Ontology	Controlled vocabulary*	genotype	http://www.geneontology.org/
PubChem	Data resource	drug	http://pubchem.ncbi.nlm.nih.gov/
RX-Norm	Controlled vocabulary	drug	http://www.nlm.nih.gov/research/umls/rxnorm/index.html
CDISC	XML schema	phenotype	http://www.cdisc.org/
ICD-10	Controlled vocabulary	phenotype	http://www.who.int/classifications/icd/
Disease Ontology	Controlled vocabulary*	phenotype	http://diseaseontology.sourceforge.net
Mammalian Phenotype	Controlled vocabulary*	phenotype	http://www.informatics.jax.org/searches/MP_form.shtml
PATO	Controlled vocabulary*	phenotype	http://obo.sourceforge.net/
ChEBI	Controlled vocabulary*	drug	http://www.ebi.ac.uk/chebi/
Pathway Ontology	Controlled vocabulary*	genotype, phenotype	http://rgd.mcw.edu/tools/ontology
SNOMED-Clinical	Controlled vocabulary	phenotype	http://www.snomed.org/snomedct/glossary.html

Table 2. List of selected resources for constructing SO-Pharm

<i>Ontology name</i>	<i>Description</i>	<i>Prefix</i>	<i>Namespace</i>
MECV	genomic variation classification	MECV	http://www.loria.fr/~coulet/ontology/mecv.owl
SNP-Ontology	genomic variations	SNPO	~/ontology/snponontology.owl
Pharmacogenetics Ontology	describes genotyping and phenotyping methods	PO	~/ontology/pharmacogeneticsontology.owl
Disease Ontology	a classification of disease	DO	~/ontology/diseaseontology.owl
Mammalian Phenotype	phenotype features	MPO	~/ontology/mammalianphenotypeontology.owl
PATO	attributes and values for phenotype description	PATO	~/ontology/pato.owl
ChEBI	molecular compounds	CHEBI	~/ontology/chebi.owl

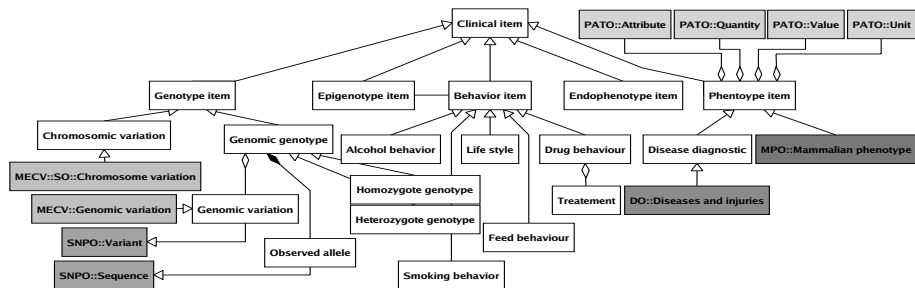


Fig. 2. UML class diagram for clinical item

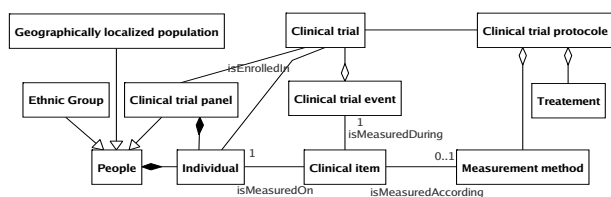


Fig. 3. UML class diagram for clinical trial

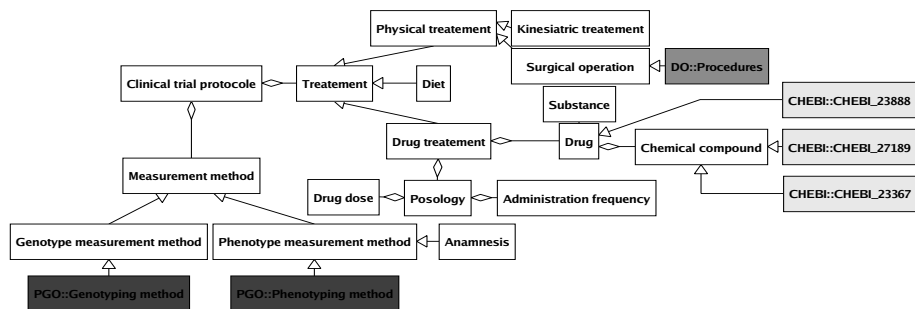


Fig. 4. UML class diagram for clinical trial protocol

according to constraints expressed in a formal language (e.g. description logics); and *terminology-based application ontologies* which are controlled vocabularies often designed to annotate biological databases [14]. Most of OBO ontologies belong to this third family -except for the PATO ontology that can be considered as a meta-ontology. In SO-Pharm, several highly specialized vocabularies such as Disease Ontology are *embedded* meaning that these ontologies are reused in an ontology having a wider scope. On the opposite, formal ontologies, such as SNP-Ontology, are high level domain representations *extending* definitions of more specific concepts pertaining from other ontologies. For example, the *variant*

concept of SNP-Ontology subsumes (i.e. extends the definition of) the *genomic variation* concept in SO-Pharm. The latter itself subsumes other more specific concepts from the OBO Sequence Ontology (e.g. *single deletion*).

In summary, SO-Pharm construction involves the design of modules favoring the reuse of concept definitions existing in other ontologies. Besides these reused concepts, additional SO-Pharm concepts and properties are defined locally.

Formalization and Implementation. SO-Pharm is implemented with the Protégé knowledge editor and coded in OWL. Formalization and implementation steps are nested. On the basis of previously designed UML class diagrams, concepts and (object and datatype) properties are formally defined in the Protégé framework. For example:

- $$\begin{aligned}
 (1) \text{ clinical_item} &\sqsubseteq \quad \exists \text{ measuredOn.individual} \\
 &\quad \sqcap \exists \text{ measuredDuring.clinical_trial_event} \\
 &\quad \sqcap \exists \text{ measuredAccording.measurement_method} \\
 \\
 (2) \text{ drug} &\sqsubseteq \quad \forall \text{ isComposedOf.chemical_compound} \\
 &\quad \sqcap \exists \text{ isPartOf.drug_treatment} \\
 &\quad \sqcap \exists \text{ isCommercialisedAs.substance}
 \end{aligned}$$

Unfortunately, no system allows an automatic conversion of UML class diagrams into OWL statements. Simple classes and associations are easily converted, but complex ones need particular attention. For example, since the description logic formalism on which OWL is based is limited to binary relationships, the translation of UML n-ary relationships is not straightforward. The most common way to represent n-ary relationships in an ontology formalism is reification [15]. In our work, conceptualization prevents n-ary relationships by preferring addition of new classes or association classes with several binary relationships.

Apart from SNP-Ontology, ChEBI and Disease Ontology which have been directly downloaded in OWL (<http://www.fruitfly.org/~cjm/obo-download/>), most external ontologies are not available in OWL. They had to be translated first. Pharmacogenetics Ontology has been manually coded in OWL from text sources. Because of redundancies, Mutation Event Controlled Vocabulary and Sequence Ontology have been manually integrated and implemented in OWL. PATO and Mammalian Phenotype Ontology have been converted from OBO format to OWL thanks to the BONG-Protégé plugin [16]. OWL-translated ontologies are then associated to namespaces and are prefixed (Table 2) for being virtually imported in SO-Pharm where they are articulated by concepts definitions:

- $$\begin{aligned}
 (3) \text{ CHEBI : molecular_entities} &\sqsubseteq \text{ chemical_compound} \\
 (4) \text{ MECV : genomic_variation} &\sqsubseteq \text{ genomic_variation} \sqsubseteq \text{ SNPO : variant}
 \end{aligned}$$

The consistency and the class hierarchy of SO-Pharm including reused ontologies have been validated with Racer 1.9 at each stage of the implementation thanks to standard reasoning mechanisms [17]. Manual construction and expert contribution appear as solid advantages for articulating existing ontologies in a sensible

way. It allows a proper use of reasoning mechanisms despite of unclear/various purpose concepts that co-exist or overlap in ontologies.

3 Preliminary Testing of SO-Pharm Semantics

As a preliminary validation of the ontology, several examples of published pharmacogenomic knowledge have been expressed with the SO-Pharm concepts. This is performed by asserting individual cases presenting genotype, treatment and phenotype features described in the literature. The assertions of individuals and related information (clinical trial, treatment) lead us to refine SO-Pharm concepts. Genotype (encompassing several genomic variation), homozygosity/heterozygosity, poor/rich metabolizer, anamnesis, treatment effect are examples of concepts added during the first round of testing in order to be able to handle the representation of selected precise pharmacogenomic examples. Groups of individuals have been artificially constituted to gather individuals presenting common traits. Three groups of individuals are presented in expression (5), (6) and (7):

- (5) *demyelinised_patient* \sqsubseteq *person*
 $\square \forall \textit{presentsGenotype}. (\exists \textit{isTheGenotypeObservedFor}. (\exists \{rs1142345\})$
 $\square \exists \textit{isComposedOf}. \exists \{G\})$
 $\square \forall \textit{presentsPhenotype}. (\forall \textit{measuredAccording}. (\exists \{6TGN_proto\})$
 $\square \forall \textit{PATO} : \textit{hasAttribute}. (\exists \{6TGN_conc\})$
 $\square \forall \textit{PATO} : \textit{hasValue}. (\exists \{high\}))$
 $\square \forall \textit{isEnrolledIn}. (\forall \textit{isDefinedBy}. (\forall \textit{isComposedOf}. (\exists \{mercaptapurine_treatment\})))$

The meaning of (5) is that demyelinised patients are persons who present both the allele G for the genomic variation rs1142345, a high concentration in 6-TGN and are treated with mercaptopurine in a clinical trial.

- (6) *over_anti_coagul_patient* \sqsubseteq *person*
 $\square \forall \textit{presentsGenotype}. (\forall \textit{isTheGenotypeObservedFor}. (\exists \{rs1057910\})$
 $\square \forall \textit{isComposedOf}. (\exists \{C\}))$
 $\square \forall \textit{isComposedOf}. (\exists \{CYP2C9.2\})$
 $\square \forall \textit{presentsPhenotype}. (\forall \textit{measuredAccording}. (\exists \{bleeding_obs\})$
 $(\forall \textit{PATO} : \textit{hasAttribute}. (\exists \{bleeding\})$
 $\square \forall \textit{PATO} : \textit{hasValue}. (\exists \{high_bleeding\})))$
 $\square \forall \textit{isEnrolledIn}. (\forall \textit{isDefinedBy}. (\forall \textit{isComposedOf}. (\exists \{warfarin_treatment\})))$

Patients with an over anti-coagulation (6) are persons who present both the allele C for the genomic variation rs1057910, the CYP2C9*2 genotype, and important bleeding and are treated with warfarin in a clinical trial.

- (7) $venous_thrombos_patient \sqsubseteq person$
 $\sqcap \forall isComposedOf. (\forall sex. (\exists \{female\}))$
 $\sqcap \forall presentsClinicalData. (\forall measuredAccording. (\exists \{drug_anamnesis\}))$
 $\sqcap \forall isComposedOf. (\exists \{oral_contraceptive\})$
 $\sqcap \forall isComposedOf. (\exists \{F2_A20210\} \sqcup \exists \{F5_A1691\})$

The patient group with venous thrombosis (7) are women who are using oral contraceptive and present the F2_A20210 or the F5_A1691 genotype. Additional assertions have led to localize and fix a few mistakes in the ontology instantiation, and to precise restrictions on object and datatype relationships. The number of required modifications decreased with each new assertion until the quasi-stability of the ontology was reached.

In view of expressing pharmacogenomic knowledge units, SO-Pharm was enriched with a simple property *mayBeRelated* that allows to link genotype item, phenotype item and chemical compound. Every required modification in the ontology is done according to a new construction iteration by updating the conceptual model, looking for reusable concepts, and finally modifying the ontology.

SO-Pharm is a crucial component for a future knowledge-based application dedicated to pharmacogenomic knowledge discovery. A complete validation has now to be conducted in the frame of the intended knowledge-based application, i.e. aimed at evaluating how SO-Pharm is able to guide the KDD process. A significant issue will be to develop appropriate wrappers to achieve heterogenous data integration as in [7].

SO-Pharm and external ontologies it includes are available (in OWL format) at <http://www.loria.fr/~coulet/ontology/sopharm.owl>. We plan to submit SO-Pharm to OBO portal to gain in visibility and facilitate further improvements.

4 Conclusion

Much of the quality of the SO-Pharm ontology relies on the initial extensive enumeration of term lists and use cases (specification and conceptualization steps). Expert interviews and overview of existing ontologies are necessary for that purpose. Interestingly case studies aimed at expressing already existing knowledge extracted from the litterature lead us to enrich SO-Pharm with additional concepts in an iterative process.

Embedding and extension strategies are used to anchor existing ontologies to SO-Pharm concepts. This conceptualization task will become more and more important since more and more autonomous ontologies are produced in the biomedical domain, e.g. for representing phenotype with formal ontologies.

References

1. Evans, W., Relling, M.: Pharmacogenomics: moving toward individualized medicine. *Nature* **429** (2004) 464–468
2. Russ B. Altman, R., Klein, T.: Challenges for biomedical informatics and pharmacogenomics. *Annu. Rev. Pharmacol. Toxicol.* **42** (2002) 113–33

3. Frawley, W., Piatetsky-Shapiro G., Matheus, C.: Knowledge Discovery in databases: An Overview, Knowledge Discovery in Databases. AAAI/MIT Press (1991) 1–30
4. Protégé ontology library [OnLine]. <http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary>
5. OBO web site [Online]. <http://obo.sourceforge.net/>
6. Oliver, D., Rubin, D., Stuart, J., et al: Ontology development for a pharmacogenetics knowledge base Pac. Symp. Biocomput. **7** (2002) 65–76
7. Coulet, A., Smail-Tabbone, M., Benlian, P. et al: SNP-Converter: An Ontology-Based Solution to Reconcile Heterogeneous SNP Descriptions. In proceedings of the 3rd Workshop on Data Integration in the Life Sciences (DILS'06), Hinxton, UK (2006)
8. Omelayenko, B.: Learning of Ontologies for the Web: the Analysis of Existent Approaches. In Proceedings of the Internat. Workshop on Web Dynamics, 8th Conference on Database Theory ICDT'01 (2001)
9. Noy, N., McGuinness, D.: Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 (2001)
10. Gabor, N.: WP3: Service ontologies and service description. DIP (Data, Information and process Integration with SW services), FP6-507483 (2005)
11. Uschold, M., King, M.: Towards a Methodology for Building Ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI-95 (1995)
12. OBO Foundry web site [OnLine]. <http://obofoundry.org/>
13. Kogut, P., Cranefield, S., Hart, L., Dutra, M., Kokar, M. Smith, J.: UML for Ontology Development. The Knowledge Engineering Review **17,1** (2002) 61–64
14. Rosse, C., Kumar, A., Mejino, J., et al: A Strategy for Improving and Integrating Biomedical Ontologies. AMIA Symposium Proceedings (2005) 639–43
15. Noy, N., Rector, A.: Defining N-ary Relations on the Semantic Web. [OnLine]. <http://www.w3.org/TR/2006/NOTE-swbp-n-aryRelations-20060412/>
16. Wroe, C., Stevens, R., Goble, C., Ashburner, M.: A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. Pac. Symp. Biocomput. **8** (2003) 624–636 Pac. Symp. Biocomput. **7** (2002) 65–76
17. Haarslev, V., Moller, R.: RACER System Description. In First Internat Joint Conference on Automated Reasoning (IJCAR'2001) 701–706 (2001)

Methodologically Designing a Hierarchically Organized Concept-Based Terminology Database to Improve Access to Biomedical Documentation^{*}

Antonio Vaquero¹, Fernando Sáenz¹, Francisco Alvarez², and Manuel de Buenaga³

¹ Universidad Complutense de Madrid, Facultad de Informática, Departamento de Sistemas Informáticos y Programación, C/ Prof. José García Santesmases, s/n, E-28040, Madrid, Spain

² Universidad Autónoma de Sinaloa, Ángel Flores y Riva Palacios, s/n, C.P 80000, Culiacán, Sinaloa, México

³ Universidad Europea de Madrid, Departamento de Sistemas Informáticos, 28670 Villaviciosa de Odón. Madrid, Spain

{vaquero, fernan}@sip.ucm.es, fjalvare@fdi.ucm.es, buenaga@uem.es

Abstract. Relational databases have been used to represent lexical knowledge since the days of machine-readable dictionaries. However, although software engineering provides a methodological framework for the construction of databases, most developing efforts focus on content, implementation and time-saving issues, and forget about the software engineering aspects of database construction. We have defined a methodology for the development of lexical resources that covers this and other aspects, by following a sound software engineering approach to formally represent knowledge. Nonetheless, the conceptual model from which it departs has some major limitations that need to be overcome. Based on a short analysis of common problems in existing lexical resources, we present an upgraded conceptual model as a first step towards the methodological development of a hierarchically organized concept-based terminology database, to improve the access to medical information as part of the SINAMED and ISIS projects.

1 Introduction

Since the days of machine-readable dictionaries (MRD), relational databases (RDB) have been a popular device to store information for linguistic purposes. Relational database technology offers many advantages, being one of its more important ones the existence of a mature software engineering database design methodology. Nevertheless, most of the efforts aimed at developing linguistic resources (LR), whether they used RDB or not, have focused on content, implementation or time-saving issues, putting aside the software engineering aspects of the construction of LR.

^{*} The research described in this paper has been partially supported by the Spanish Ministry of Education and Science and the European Union from the European Regional Development Fund (ERDF) - (TIN2005-08988-C02-01 and TIN2005-08988-C02-02).

Many authors use the term “software engineering” synonymously with “systems analysis and design” and other titles, but the underlying point is that any information system requires some process to develop it correctly. The basic idea is that to build software correctly, a series of steps (or phases) are required. These steps ensure that a process of thinking precedes action: thinking through “what is needed” precedes “what is written”. Although software engineering spans a wide range of problems, we will focus here on the database design aspects.

As it will be seen later, design issues are important when using RDB. Moreover, as we stated in [1], design is also important because in order to develop, reuse and integrate diverse available resources, into a common information system, perhaps distributed, requires compatible software architectures and sound data management from the different databases to be integrated. With that in mind, we have defined a methodology [1], for the design and implementation of ontology-based LR using RDB and a sound software engineering approach. Nevertheless, the conceptual model we propose as a point of departure of the methodology has some major limitations, which have to be overcome in order to create structurally sound LR.

In this paper, we will focus on the ontology representation limitations of our previous model (leaving the lexical side limitations for a future paper), and create a conceptual model of the ontological part, that overcomes such limitations as part of our efforts to have a solid foundation for action. Our final goal is to create a LR (a hierarchically organized concept-based terminology database) that will be part of an intelligent information access system that integrates text categorization and summarization, to improve information access to patient clinical records and related scientific documentation, as part of the SINAMED and ISIS projects [2].

The rest of the paper is organized as follows. In section 2, the advantages and disadvantages of RDB are pointed out, as well as the importance of database design in the construction of ontology-based LR. In section 3, some common problems of LR are summarized, and the need to develop methodologically engineered application-oriented LR is signaled. In section 4, the methodological gaps of past developing efforts are underlined. In section 5, a set of ideas intended to help developers to formally specify and clarify the meaning of concepts and relations are depicted. In section 6, a conceptual model that integrates the aforementioned ideas is introduced and described. Finally, in section 7 some conclusions and future work are outlined.

2 Designing LR Using RDB

RDB present a series of advantages that have been taken into account when used to construct databases for linguistic purposes [1, 3, 4, 5, 6]. From a software engineering point of view, their main advantage is that they provide a mature design methodology, which encompasses several design stages that help designing consistent (from an integrity point of view) databases. This methodology comprises the design of the conceptual scheme (using the Entity/Relationship (E/R) model), the logical scheme (using the relational model), and the physical scheme.

However, RDB have various drawbacks when compared to newer data models (e.g., the object-oriented model): a) Impossibility of representing knowledge in form of rules; b) Inexistence of property inheritance mechanisms; and c) Lack of expressive power to represent hierarchies. In spite of this, by following a software engineering approach, that is, by paying attention to the database design issues [4], most of these drawbacks can be overcome, and thus, let us take advantage of all the benefits of RDB.

For instance, in [5] we can see how an UML (object-oriented) model is implemented within a RDB in a way that supports inheritance and hierarchy. Another similar example is found in [7], where the authors reproduce the structure of the Mikrokosmos ontology, using the E-R model. Other models [3, 8], although machine translation oriented follow a purely linguistic approach, and are not intended to overcome any of the limitations of the relational data model.

As it can be deduced, we have focused on the limitations of RDB to represent ontologies. There are several reasons why we have done that. First, our work is focused on the design and implementation of ontology-based LR using RDB [1]. Second, it has been proved by [9] that the use of a hierarchically organized concept-based terminology database, improves the results of queries on clinical data, and such is the goal of our projects. Third, we agree with [4, 10, 11, 12], when they state that the computationally proven ontological model, with two separated but linked levels of representation (i.e. the conceptual-semantic level and the lexical-semantic level) is our best choice for linguistic knowledge representation.

We have only found one reference, of a development effort that follows our software engineering approach for the development of ontology-based LR: the aforementioned work of [7]. The difference between our model [1] and the one in [7] is that ours only follows the ontological semantics ideas of Mikrokosmos; it does not recreate its frame-based structure. Nevertheless, although the model in [7] replicates the powerful ontological structure of Mikrokosmos in a RDB, it inherits all its problems (some will be described in the next section). As for the model we present in [1], it has a thesaurus-like structure where the concepts of the ontology are linked by a single implicit and imprecise relation; a situation that is problematic and severely limits the model, as it will be shown next.

3 Some Common Problems in LR

It is relatively easy to create a conceptual model of a LR. As seen in the previous section, this has already been done. However, existing LR (ontology-based or not) are plagued with flaws that severely limit their reuse and negatively impact the quality of results. Thus, it is fundamental to identify these flaws in order to avoid past and present mistakes, and create a sound conceptual model that leads to a LR where some of these errors can be avoided.

Most of the problems of past and present LR have to do with their taxonomic structure. For instance, once a hierarchy is obtained from a Machine-Readable

Dictionary (MRD), it is noticed that it contains circular definitions yielding hierarchies containing loops, which are not usable in knowledge bases (KB), and ruptures in knowledge representation (e.g., a utensil is a container) that lead to wrong inferences [13]. WordNet and Mikrokosmos have also well-known problems in their taxonomic structure due to the overload of the is-a relation [14, 15]. In addition, Mikrokosmos represents semantic relations as nodes of the ontology. This entails that such representation approach where relations are embedded as nodes of the ontology is prone to suffer the same is-a overloading problems described in [14, 15], as well as the well-known multiple inheritance ones (figure 1 illustrates this point by showing part of the Mikrokosmos ontology). In the biomedical domain, the UMLS has circularities in the structure of its Metathesaurus [16], because of its omnivorous policy for integrating hierarchies from diverse controlled medical vocabularies whose hierarchies were built using implicit and imprecise relations. Some of the consequences of these flaws, as well as additional ones have been extensively documented in [10, 11, 14, 17, 18, 19, 20, 21] for these and other main LR.

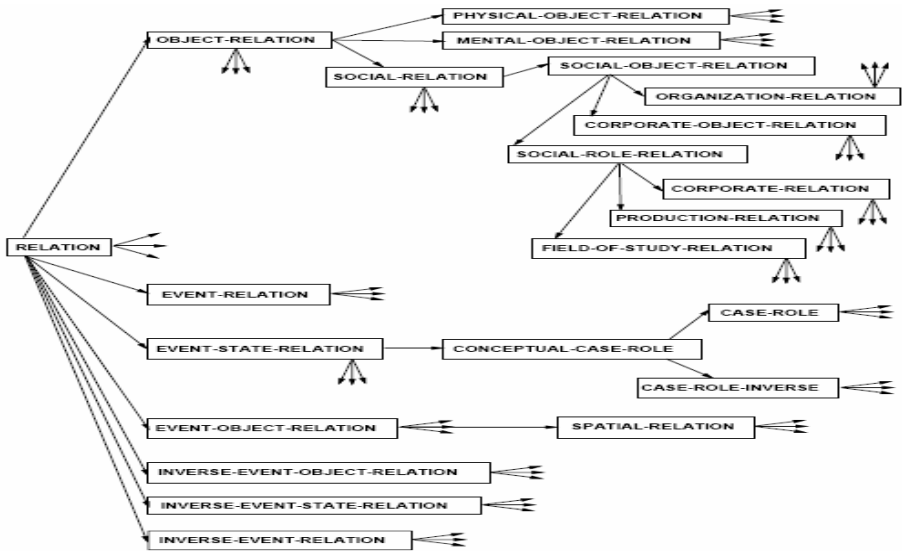


Fig. 1. Embedded Relations in the Mikrokosmos Ontology

3.1 Methodologically Engineered Application-Oriented LR

We have come a long way from the days of MRD. However, still today, the focus is on coverage and time-saving issues, rather than on semantic cleanness and application usefulness. Proof of this are the current different merging and integration efforts aimed at producing wide-coverage general LR [21, 22], and the ones aimed at (semi)automatically constructing them with machine learning methods [23, 24]. However, no amount of broad coverage will help raise the quality of output, if the

coverage is prone to error [11]. We should have learned by now that there are no short cuts, and that most experiments aimed at saving time (e.g., automatically merging LR that cover the same domains, or applying resources to NLP that are not built for it, like machine-readable dictionaries and psycholinguistic-oriented word nets) are of limited practical value [25]. Furthermore, in the current trend of LR development issues such as how to design LR are apparently less urgent, and this is haphazard. More attention must be paid on how LR are designed and developed, rather than what LR are produced.

The experience gained from past and present efforts clearly points out that a different direction must be taken. As [18] pointed out back in the days of MRD: “rather than aiming to produce near universal LR, developers must produce application-specific LR, on a case by case basis”. In addition, we claim that these LR must be carefully conceived and designed in a systematic way, according to the principles of a software engineering methodology. This is especially true if RDB are to be used as a knowledge representation schema for LR.

4 Methodological Gaps in the Development of LR Using RDB

Since we are interested in the development of a LR using RDB, it is worth mentioning that all the cited efforts in section 2, although they produced useful resources, they forgot about the methodological nature of RDB. They all stopped at the conceptual design stage. Thus, there is not a complete description of the entities, relationships and constraints involved in the conceptual and logical design of the DB.

The methodology we propose in [1] encompasses all of the database design phases. Nonetheless, the conceptual model from which it departs has several problems with respect to ontology representation; mainly, its does not foresee any control and verification mechanism for clarifying the semantics of relations, a problem that as seen in section 3 is of main concern.

Hence, if we are to design a hierarchically organized concept-based terminology database using RDB, our conceptual model must take also into account the semantic relations issue. As a first step, we enhance the conceptual model presented in [1] as shown in the next section.

5 Refining the Semantics of Concepts and Relations

In order to give our first step towards the enhancement of the conceptual model, we need to clearly state what are the elements that will be abstracted and represented in our upgraded conceptual model, that will help us to: a) build application-oriented LR (as pointed out in section 3.1); and b) avoid the problems present in existing LR as described in section 3.

These elements are concepts, properties of concepts, relations, and algebraic and intrinsic properties of relations. They will help an ontology developer to specify for concepts and relations formal and informal semantics that clarify the intended meaning of both entities in order to avoid the problems discussed in section 3.

Informal semantics are the textual definitions for both concepts and relations, as opposed to formal semantics that are represented by the properties of concepts and relations.

However, the fact that these elements will be part of the enhanced conceptual model does not imply that they are an imposition but rather a possibility, a recommendation that is given to each ontology developer. In the following, we detail the elements surrounding the basic element of our model: concepts.

5.1 Properties of Concepts

These are formal semantic specifications of those aspects that are of interest to the ontology developer. In particular, these specifications may be the metaproperties of [15] (e.g., R, I, etc.) In our application-oriented approach to LR development, only the properties needed for a concrete application domain should be represented. These properties play an important role in the control of relations as it will be seen later.

5.2 Relations

Instead of relations with an unclear meaning (e.g. subsumption), we propose the use of relations with well-defined semantics, up to the granularity needed by the ontology developer. Moreover, we refuse to embed relations as nodes of the ontology (because of the problems commented in section 3) or to implicitly represent any relation as it is done in Mikrokosmos with the is-a relation. We call these, explicit relations. This represents a novelty and an improvement when compared to similar design and implementation efforts as [7] based on RDB. In the next two subsections, we will describe the elements that help clarifying the semantics of relations.

5.3 Algebraic Properties of Relations

The meaning of each relation between two concepts must be established, supported by a set of algebraic properties from which, formal definitions could be obtained (e.g., transitivity, asymmetry, reflexivity, etc.). This will allow reasoning applications to automatically derive information from the resource, or detect errors in the ontology [26]. Moreover, the definitions and algebraic properties will ensure that the corresponding and probably general-purpose relational expressions are used in a uniform way [26]. Tables 1 and 2 (taken from [26]) show a set of relations with their definitions and algebraic properties.

Table 1. Definitions and Examples of Relations

Relations	Definitions	Examples
$C \text{ is-a } C_1$	Every C at any time is at the same time a C_1	<i>myelin is-a lipoprotein</i>
$C \text{ part-of } C_1$	Every C at any time is part of some C_1 at the same time	<i>nucleoplasm part-of nucleus</i>

Table 2. Algebraic Properties of Some Relations

Relations	Transitive	Symmetric	Reflexive
Is-a	+	-	+
part-of	+	-	+

5.4 Intrinsic Properties of Relations

How do we assess, for a given domain, if a specific relation can exist between two concepts? The definitions and algebraic properties of relations, although useful are not enough. As [15] point out, we need something more. Thus, for each relation, there must be a set of properties that both a child and its parent concept must fulfill for a specific relation to exist between them. We call these properties, intrinsic properties of relations. For instance, in [15] the authors give several examples (according to their methodology) of the properties that two concepts must have so that between them there can be an is-a relation.

6 Designing the Conceptual-Semantic Level of the Concept-Based Terminology Database

In this section, we present the conceptual model (an E/R scheme upgraded from our model in [1]) shown in figure 2, for the conceptual-semantic level of our future terminology database as a result of the first design phase, where all the ideas described in section 5 have been incorporated. However, as it was previously established, the model will reflect only the ontology part of our future hierarchically organized concept-based terminology database.

The entity set Concepts denotes the meaning of words, and it has two attributes: ConceptID (artificial attribute intended only for entity identification), and ConceptDefinition, intended for the textual definition of the meaning (informal semantics). The entity set ConceptProperties represents the set of formal properties described in section 5.1, and it has one attribute: ConceptProperty used to represent each property.

The entity set Relations represents the set of relations that can exist in an ontology, and it has two attributes: Relation that captures the textual name of each relation (e.g., is-a, part-of, etc.), and RelationDefinition for the textual definition of relations (informal semantics) as illustrated in table 1.

The entity set AlgebraicProperties represents the properties of relations (formal semantics) as seen in table 2, and it has one attribute: AlgebraicProperty that denotes each algebraic property. The entity set IntrinsicProperties conveys the set of properties mentioned in section 5.4 and has one attribute: IntrinsicProperty which represents each intrinsic property.

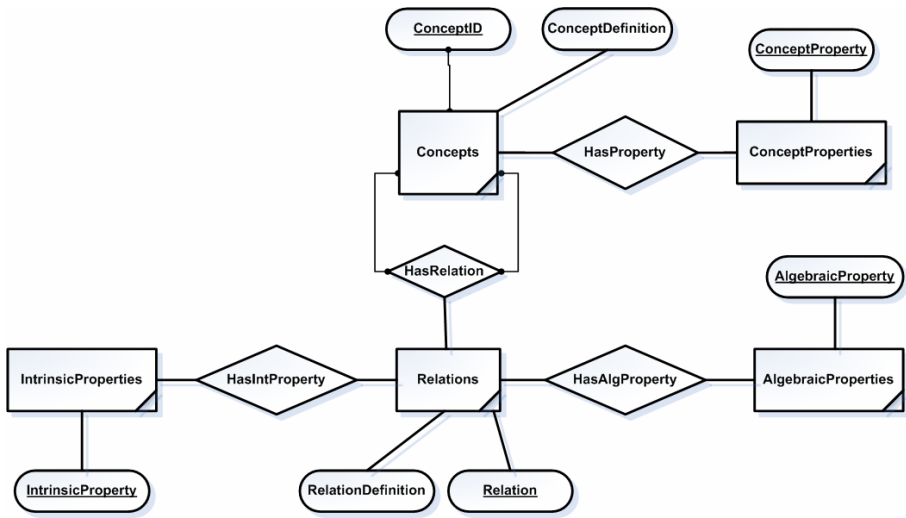


Fig. 2. Conceptual Model for an Ontology-Based LR

The relationship set *HasProperty* is used to assign properties to concepts. The ternary relationship set *HasRelation* is used to represent that two concepts in an ontology can be linked by a given relation. The relationship set *HasAlgProperty* is used to convey that relations could have attached a set of algebraic properties; the same applies for the relationship set *HasIntProperty*, but for intrinsic properties.

7 Conclusions and Future Work

The use of RDB to represent lexical knowledge provides a complete software engineering methodological approach for the design of the database that will contain the LR. However, the approaches that use this technology sometimes only present an E-R schema and forget about the rest of the DB development stages or simply state that they use RDB. This is far from being adequate, as LR to be used by domain specific applications need to be developed in such a way that all the modeling choices are clearly stated and documented.

With that in mind, we have chosen to develop our future terminology database following a sound software engineering methodology. However, the proposed conceptual model of the methodology had some major limitations. In order to overcome them, we modified it based on an analysis of common problems in LR. The new model can now account for any number of ontological relationships (as long as they are binary), and we have incorporated a set of ideas that help designing application-oriented LR where the semantics of relations is clearly stated and the use of relations can be controlled (e.g., the model allows the integration of the OntoClean [15] method for evaluating taxonomies). Moreover, although we have selected RDB

to represent lexical and conceptual knowledge, the model is totally independent of any knowledge representation schema (i.e., databases or knowledge bases).

We still have to go through the logical and physical design stages of the database. However, we have taken a first step towards our final goal, by clearly stating and depicting the structure, scope and limitations of our future LR. Moreover, we have focused on the ontology side of the model; however, the lexical side of our previous model (see [1]) also needs to be upgraded as it is quite limited. Thus, we are considering the integration of the E-R model for the lexical side of an ontology-based LR proposed and described in [4].

A thing that must be clearly understood is that our efforts lean towards the establishment of a software engineering methodology for the design and implementation of ontology-based LR using RDB. However, it is not a methodology aimed at saving time by: a) constructing or extracting a LR from texts using machine learning methods [23, 24] or b) merging different LR into a definitive one [21, 22]. We follow a software engineering approach (where thinking precedes action) by focusing on analysis, design and reuse (as understood by software engineering) aspects. Thus, we apply the principled methods and techniques of software engineering (which guide the development of user-oriented, readable, modular, extensible, and reusable software) to the design and implementation of ontology-based LR.

Finally, a very important aspect in developing a LR is the availability of software tools for its enlargement and modification. However, the majority of the management software tools for LR are just briefly described, by pointing out their features, and although some are extensively described [7, 17], there is no declared software engineering approach for their development [1]. Although not covered in this paper, our methodology takes also into account this important aspect.

References

1. Sáenz, F. and Vaquero, A. 2005. Applying Relational Database Development Methodologies to the Design of Lexical Databases. Database Systems 2005, IADIS Virtual Multi Conference on Computer Science and Information Systems (MCCSIS), ISBN 972-8939-00-0, (2005)
2. Maña, M., Mata, J., Domínguez, J.L, Vaquero, A., Alvarez, F., Gomez, J., Gachet, D., De Buenaga, M. Los proyectos SINAMED e ISIS: Mejoras en el Acceso a la Información Biomédica Mediante la Integración de Generación de Resúmenes, Categorización Automática de Textos y Ontologías. En Actas del XXII Congreso de la Sociedad Española de Procesamiento del Lenguaje (SEPLN), (2006)
3. Bläser, B; Schwall, U and Storrer, A. Reusable Lexical Database Tool for Machine Translation. In Proceedings of the International Conference on Computational Linguistics -- COLING'92, volume II, (1992) pp. 510-516.
4. Moreno A. Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática. Estudios de Lingüística Española, vol(9). (2000)
5. Hayashi, L. S. and Hatton, J. Combining UML, XML and Relational Database Technologies - The Best of all Worlds for Robust Linguistic Databases. In Proceedings of the IRCS Workshop on Linguistic Databases. (2001).

6. Wittenburg, P., Broeder, D., Piepenbrock, R., Veer, K. van der. Databases for Linguistic Purposes: a case study of being always too early and too late. In Proceedings of the EMELD Workshop. (2004).
7. Moreno, A. and Pérez, C. Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases. In Proc. of the 2nd International Conference on Language Resources and Evaluation, (2000) pp 1061-1067.
8. Tiedemann, J. MatsLex: A multilingual lexical database for machine translation. In Proc. of the 3rd International Conference on Language Resources and Evaluation, (2002), pp 1909-1912.
9. Lieberman, M. The Use of SNOMED to Enhance Querying of a Clinical Data Warehouse. A thesis presented to the Division of Medical Informatics and Outcomes Research and the Oregon Health & Sciences University School of Medicine in partial fulfillment of the requirements for the degree of Master of Science. (2003)
10. Nirenburg, S., McShane, M. and Beale, S. The Rationale for Building Resources Expressly for NLP. In Proc. of the 4th International Conference on Language Resources and Evaluation, (2004).
11. McShane, M.; Nirenburg, S. and Beale, S. An implemented, integrative approach to ontology-based NLP and interlingua . Working Paper #06-05, Institute for Language and Information Technologies, University of Maryland Baltimore County, (2005)
12. Cimino, J. Desiderata for Controlled Medical Vocabularies in the Twenty-first Century. *Methods of Information in Medicine*, 37(4-5):394-403, (1998)
13. Ide, N., and Veronis, J. Extracting Knowledge Bases from Machine-Readable Dictionaries: Have we wasted our time? In Proc. of the First International Conference on Building and Sharing of Very Large-Scale Knowledge Bases, (1993)
14. Guarino, N. Some Ontological Principles for Designing Upper Level Lexical Resources. A. Rubio et al. (eds.), In Proc. of the First International Conference on Language Resources and Evaluation, (1998) pp 527-534.
15. Welty, C. and Guarino, N. Supporting ontological analysis of taxonomic relationships", *Data and Knowledge Engineering* vol. 39(1), (2001) pp. 51-74.
16. Bodenreider O. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. In Proceedings of the AMIA Symposium, (2001)
17. Feliu, J.; Vivaldi, J.; Cabré, M.T. Ontologies: a review. Working Paper, 34. Barcelona: Institut Universitari de Lingüística Aplicada. DL: 23.735-2002 (WP), (2002)
18. Evans, R., and Kilgarriff, A. MRDs, Standards and How to do Lexical Engineering. Proc. of 2nd Language Engineering Convention, (1995) pp. 125–32.
19. Burgun, A. and Bodenreider, O. Aspects of the Taxonomic Relation in the Biomedical Domain. In Proc. of the 2nd International Conference on Formal Ontologies in Information Systems, (2001)
20. Martin, P. Correction and Extension of WordNet 1.7. In Proc. of the 11th International Conference on Conceptual Structures, (2003) pp 160-173.
21. Oltramari, A.; Prevot, L.; Borgo, S. Theoretical and Practical Aspects of Interfacing Ontologies and Lexical Resources. In Proc. of the 2nd Italian SWAP workshop, (2005).
22. Philpot, A., Hovy, E. and Pantel, P. The Omega Ontology. 2005. In IJCNLP Workshop on Ontologies and Lexical Resources, (2005) pp. 59-66.
23. Makagonov, P., Ruiz Figueroa, A., Sboychakov, K. and Gelbukh, A. Learning a Domain Ontology from Hierarchically Structured Texts. In Proc. of Workshop "Learning and Extending Lexical Ontologies by using Machine Learning Methods" at 22nd International Conference on Machine Learning, (2005)

24. Makagonov, P., Ruiz Figueroa, A., Sboychakov, K. and Gelbukh, A. Studying Evolution of a Branch of Knowledge by Constructing and Analyzing Its Ontology. In Christian Kop, Günther Fliedl, Heinrich C. Mayr, Elisabeth Métais (eds.). Natural Language Processing and Information Systems. 11th International Conference on Applications of Natural Language to Information Systems, (2006).
25. Nirenburg, S., McShane, M., Zabudowski, M., Beale, S. and Pfeifer, C. Ontological Semantic text processing in the biomedical domain. Working Paper #03-05, Institute for Language and Information Technologies, University of Maryland Baltimore County, (2005)
26. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies. *Genome Biology*, 6(5), (2005)

A Proposal for a Gene Functions Wiki

Robert Hoehndorf^{1,2,3}, Kay Prüfer², Michael Backhaus^{1,2}, Heinrich Herre^{1,3},
Janet Kelso², Frank Loebe^{1,3}, and Johann Visagie²

¹ Research Group Ontologies in Medicine, Institute for Medical Informatics, Statistics and
Epidemiology, University of Leipzig

² Department of Evolutionary Genetics, Max-Planck-Institute for Evolutionary Anthropology

³ Institute for Informatics, University of Leipzig

Abstract. Large knowledge bases integrating different domains can provide a foundation for new applications in biology such as data mining or automated reasoning. The traditional approach to the construction of such knowledge bases is manual and therefore extremely time consuming. The ubiquity of the internet now makes large-scale community collaboration for the construction of knowledge bases, such as the successful online encyclopedia “Wikipedia”, possible.

We propose an extension of this model to the collaborative annotation of molecular data. We argue that a semantic wiki provides the functionality required for this project since this can capitalize on the existing representations in biological ontologies. We discuss the use of a different relationship model than the one provided by RDF and OWL to represent the semantic data. We argue that this leads to a more intuitive and correct way to enter semantic content in the wiki. Furthermore, we show how formal ontologies could be used to increase the usability of the software through type-checking and automatic reasoning.

1 Background

Recent technology developments have lead to the availability of genome sequence and annotation data for a wide variety of species. More than twelve mammalian and thousands of non-mammalian genomes have been sequenced and are publicly available. Large volumes of biological data including sequences, structures, functions, pathways and networks are now available. One of the major challenges in the field of bioinformatics is to store and represent this data in a way which enables researchers to analyze data integrated from diverse domains[1–3].

Understanding the relationship between phenotype (identifiable traits) and genotype (heritable information), and the influence of environmental factors on both, remains a major research area. The interaction between various factors resulting in the phenotype is highly complex and requires a detailed understanding of multiple areas of biology. Among these, gene function is considered central.

One approach to solve the problem of representing and structuring data about genes and gene products is the Gene Ontology[4]. In its current release, the Gene Ontology has more than 19000 concepts. Each concept names either a molecular function, or a biological process to which a gene product is associated, or the location of the gene product within the cell. The concepts are linked by two relations, part-of and is-a. Both

relations satisfy the transitivity property (the statement, that if some gene is associated with a category it is also associated with all the linked super-categories)¹. The Gene Ontology is maintained by a team of curators and relies heavily on input from the community for its content and correctness. Linking gene products to the GO is performed independently for various databases and model organisms. While the GO provides information about genes that have certain functions, or which are known to be in some way associated with a function, process or cellular component, it does not provide a review of the research and discussions leading to this hypothesis – information which is crucial in the analysis of the function of a gene product and in discussing it in context with previous assumptions.

While the current biological databases and ontologies provide information about many features of a gene, the discussion of hypotheses regarding the biological function is missing. Wang [5] raised the question whether this problem could be solved by a wiki similar to wikipedia[6]. In his opinion the striking advantage of a wiki is the implicit community involvement. He concludes that “[A] wiki on gene function, which utilizes the collective brain power of biologists around the world, would be an invaluable tool for biological sciences.”

2 The Application of Wikis and the Need for Structure

Wang’s idea to use a wiki for the collaborative work on a comprehensive description of the function of genes is inspired by the success story of the online encyclopedia Wikipedia. A recent comparison of the traditional to the Wikipedia way of gathering information shows that this approach is competitive [7]. While these results are encouraging and show that the idea has potential, we see one major obstacle to this approach: the users of Wikipedia use the encyclopedia in a more or less traditional way to find information about one keyword (and relevant further articles by following links) while large scale analyzes in biology require the extraction of information regarding many items at the same time in a format suitable for computation. We propose the use of an extended semantic wiki as suitable for addressing these issues and providing data in biologically relevant formats. In this semantic wiki, instances or concepts (which are instances of some meta-category “concept”) are treated as wiki pages and relationships between them are treated as hyperlinks between wiki pages.

Furthermore, since ontologies play a major role in the description of data about genes, the structure of existing ontologies must be compatible with the wiki. Therefore, a semantic wiki will allow for the description of information about transcripts and for the collaborative development of a biomedical knowledge base which is used for describing transcripts and other biological entities. In this paper, we will call this wiki the “gene function wiki”.

2.1 Semantic Wikis

In order to enable the users of the gene function wiki to extract the contents in a machine readable format, the syntax of the standard wiki page has to be enhanced. Many wikis

¹ In the terminology of the Gene Ontology this is called the “True-Path-Rule”.

which use a formal model to represent content have been developed in recent years. Most of these semantic wikis use Semantic Web technology (OWL and RDF) as their underlying representation formalism. For our purposes, semantic wikis can be divided into two main categories: ontology editors with features to support collaboration or transactions[8,9], and wikis which are extended to allow for the semantic annotation of links or attributes[10–13]. The first often support more expressive constructs, such as OWL-DL, and are intended for users who have experience in the creation and use of ontologies and knowledge bases and are interested in the collaborative features, while the latter support RDF and rarely more expressive formalisms, and tend to be focused on users with a main interest in adding semantic context to the edited text.

The gene function wiki is intended for use by biologists with limited knowledge of formal logic, the Semantic Web or ontologies. It is therefore vital to keep the front-end intuitive, while representing the complexity of interactions between genes. Intuitive and commonsense ways for knowledge acquisition are of major importance if the application is to be widely adopted.

2.2 Requirements

We summarize here the requirements for building a gene function wiki. First, it must be possible to use and represent the structure of ontologies in the wiki in a way which can be queried rapidly. For example, a query for all the genes which are involved in apoptosis in neural crest cells relies heavily on the structure and semantics of the relations in the Gene Ontology[4] and the Celltype Ontology[14], and requires a structured representation of the information in our wiki describing gene functions.

A different type of formal information is necessary to answer questions about the exact way in which a gene is taking part in a biological process, i.e., whether the gene product just supports the function, or if it is an integral part of the chemical reaction. We need to be able to distinguish between a gene product which participates in a process from one which results from it. Ontologies such as the Ontology of Functions[15,16] require n -ary relations, so queries for the exact role a gene product plays in a relation are relevant. We may want to restrict these queries even further, for example by requesting a specific author or evidence. All this requires a way to add structured information to the wiki and well-defined semantics for the relations and attributes used. Furthermore, access to the relations used in biomedical ontologies and their semantics are needed to represent and query the description of gene functions.

Because a major part of the knowledge which is developed is represented as text, the use of a simple frontend is essential to the application. We will therefore extend a semantic wiki to be adapted specifically to the problem at hand, since none of the existing prototypes satisfies our needs.

In addition to the use of existing biomedical ontologies in the description of the functions of genes and gene products, new concepts which are not yet part of any biomedical knowledge base may be required to describe the functionality of a gene, and it must be possible to add them to the wiki and interrelate them with the existing biomedical ontologies. Therefore, another application for the gene function wiki will be the use as a collaborative ontology and knowledge base curation system[17]. For this task, it is

crucial to provide intuitive ways to enter semantic content in the wiki, and to implement automatic checks for logical errors.

Finally, it must be possible to describe concepts as well as individuals in the gene function wiki. Concepts are used to categorize findings in an experiment, which is an individual. Also, annotating some experiment requires the description of individuals, while the conclusions are often abstractions and generalizations, therefore concepts.

3 Representation Language

Most semantic wikis allow only for the representation of binary relations, due to the restriction of the RDF format to binary relations. However, many relations in biomedicine, such as the annotation relation, can take more arguments. What is needed is an intuitive way to model n -ary relations in a semantic wiki. This can be done by keeping the original understanding of the semantic relations, as a typed link to another page in the wiki, but adding argument slots to the relation, which may be filled by arguments of an appropriate type (further discussed later on).

3.1 OWL and RDF

OWL and RDF are specifications for a metadata model maintained by the World Wide Web Consortium (W3C). RDF allows one to make statements about things in the form of subject – predicate – object triples, where the subject is the resource which is described, the relation represents a specific aspect of this resource, and the object is the value of the relationship.

OWL is an extension of RDF, and can be used to share and publish ontologies. OWL comes in three flavors, OWL-Lite, OWL-DL and OWL-Full. The expressive power of OWL-DL is equivalent to the description logic $SHOIN(D)$, and the expressive power of OWL-Lite to $SHIF(D)$ [18].

It is possible to reify statements made in RDF, and treat them as a new resource. This can be used to introduce n -ary relations. It is further possible to introduce a concept in OWL that takes n attributes as an n -ary relation in OWL. This can be used to export our data model, which is described later, to OWL.

3.2 Relations and Roles

K. Devlin [19] describes a model for relationships which is close to the everyday use of relations in, for example, natural language expressions, and which meets our requirements. Relations in [19] are specified by means of a *name* and named *argument roles*, which are slots in which objects of a specified *type* can be placed. It is possible to omit arguments in a use of the relation. However, a minimality condition is defined for each relation, defining which argument slots must be filled in order for a relation to be meaningful. For example, the relation *partOf* could be described as $\langle \text{partOf} | \text{part, whole, context} \rangle$ or the relation *eats* as $\langle \text{eats} | \text{eater, eatenObject, means, location, time} \rangle$ where *means* denotes the means used to eat (such as a knife). The statement “John eats an apple now” would be represented as $\ll \text{eats, eater} \rightsquigarrow \text{John, eatenObject} \rightsquigarrow \text{apple,}$

$time \rightsquigarrow t_{now} \gg$. Note that the *means* and *location* argument roles are unfilled. A minimality condition here could state that either the *eater* or the *objectEaten* must be filled. These argument roles will also restrict the type of object which may fill the role, as we will discuss in section 4.

RDF can be integrated in this view: for some RDF triple *subject, predicate, object*, we can define the relation $\langle relation, subject, object \rangle$. The relationship model we use is similar to UML associations[20] and topic maps[21, 22]. Our representation of the structure of relations is also close to the account on roles given in [23, 24]. We will show this similarity in the discussion of our data-model in section 5.

4 Ontological Type System

Part of the strategy which lead to the success of wikis is that they leave their users a maximum of liberty; there is no structure in a wiki except for the one provided by the users of the wiki. We, however, want to use a wiki for the creation of a structured knowledge base, in a domain in which rich representations of structures exist in the form of ontologies. In this section, we address the question how to add such a structure without limiting the ability a user has to edit information in the gene function wiki. Instead we wish to provide an easy reference to the structure which is available in the wiki itself, and to structures which have been developed outside the wiki and will be used to annotate content of the wiki, such as the Gene Ontology.

But first, let us collect some examples of what kind of structure we talk about. Biological processes in the Gene Ontology, for example, are related using two relations, *part-of* and *is-a*. Additionally, conceptualizations and formalizations of the most general entities in biology are developed[25, 26]. Some of these conceptualizations are new and still need to gain wide acceptance in the biological community[15], but others such as the need for the concepts of “function”, “process” and “localization” as included in the Gene Ontology are accepted throughout the scientific biological community. At least the terminology and structure that the Gene Ontology provides must be usable for the description of the gene products.

We provide a structural layer of the wiki in the form of a biomedical core ontology. This core ontology gives natural language and formal definitions of the most general biomedical concepts, such as *biological process*, *biological function*, or *organism*. Additionally, the core ontology defines relations between these concepts, for example a relation *Realizes* between processes, functions, and objects. Furthermore, it defines the upper categories of all the biomedical domain-ontologies which are used in conjunction with the wiki.

Since a core ontology is a rigorous yet abstract formalization of the entities and relations of a domain, all of the (semantic) information in the wiki can be embedded in the core ontology. Making a general set of relations available leads to less redundancy in the definition of new relations. For example, the relation *is-a* could also be named *subclass-of*, *specialization-of*, or *subsumed-by*. By providing one relation in the core ontology, all these names will be derivatives of this one relation, which has been formally defined.

Furthermore, the core ontology can be used as a type system for the relations and concepts in the gene function wiki as described in [27]. For example, a relation *has-function* could be specified as relating only biological functions, biological objects, and a context. With concepts in the gene function wiki that have types, two things can be done whenever some concepts are related using the *has-function* relation: (1) verifying whether the arguments of this relation are of the type which is specified by the *has-function* relation, and (2) automatically classifying the arguments of the relation as arguments of the appropriate type.

As an example of what can be done, assume that a *hasCellFunction* relation was defined in the following way, where the arguments are specified as (*role, type*) tuples, and *cell*, *bioFunction*, and *situation* are concepts defined in the core ontology:

$\langle \textit{hasCellFunction}, (\textit{bearer}, \textit{cell}), (\textit{function}, \textit{bioFunction}), (\textit{context}, \textit{situation}) \rangle$

A concept (wiki page) *A* which occurs in the *bearer* role of this relation is automatically classified as a cell, and inherits all properties of cells that are defined in the core ontology (such as the potential to be part of some tissue) and the ontologies which are embedded in the core ontology (such as the existence of a part which is a membrane as defined in the Gene Ontology). If the same concept *A* occurs in a different relation filling a role which is typed as *organism*, *A* is reclassified as $\textit{cell} \sqcap \textit{organism}$, a mono-cellular organism (e.g., a bacteria).

Also, if cells in the core ontology were defined as a superclass of things which have as part a nucleus, and the concept *A* in the wiki had a nucleus as part, then *A* can be classified as cell.

This type system can also be used for information retrieval: using the core ontology, it is possible to query for all the processes to which a gene is associated.

We use the core ontology GFO-Bio[26] which is based on the top-level ontology General Formal Ontology[28] (GFO). This is particular useful because the GFO includes a well-developed analysis of ontological categories such as universals, concepts and symbols. Symbols play an important role in the description of genes, DNA or RNA. Although the GFO is a top-level ontology which is formalized in first order logic, an OWL-DL version is available for conceptual modeling purposes. GFO-Bio, which is based on the GFO, is available in OWL-DL as well.

5 Data-Model

In addition to storage for the text in the wiki articles, we need an additional place to store the semantic data, similar to other semantic wikis[10].

A UML diagram of the data model is shown in figure 1. What can be seen is the distinction between concepts on the left hand side, and their instances – individuals – on the other. Relations can be defined as concepts by a relation name and a number of role–type pairs. The type of a role can be a disjunctive type, which is the reason for the $n : m$ relationship between roles and types. Roles are bound to one relation and cannot be reused.

On the other side are instances. Instances of relations – called *relators* in the General Formal Ontology (GFO) [28] – can be split into instances of roles, called

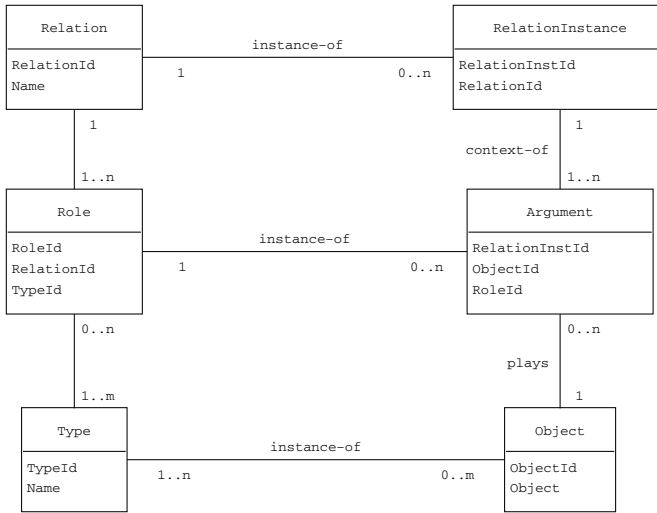


Fig. 1. UML Diagram of the data model

qua-individuals in [29], which are played by objects. In the terminology of [24, 23], roles – here represented in the argument table – are dependent on a context (the instance of the relation, or the relator) and a player (the object). These constraints are represented as restrictions on the cardinality of the relation arguments have to objects and relators.

6 Discussion and Future Research

There are still many problems which require a solution. Although there are some biomedical core ontologies already available, none of them has ever been applied as discussed in our proposal. It is to be expected that in addition to the concepts of the biomedical domain, concepts which are not part of this domain must be added. Types such as numbers, coordinates, strings, author names, et cetera are not a part of the current core ontologies for biomedicine, but will be required for application in the wiki.

Another question is how far the idea we discussed here could be generalized to other domains. It is tempting to use the same architecture for other domains such as chemistry, physics or even social sciences, by adding different core ontologies for other domains, and embed them into the same top-level ontology, GFO. However, it remains open, whether interesting parts of other domains can be formalized with the language we provide, or if richer formalisms are required. For example, the core ontology of mathematics is set theory. But formalizing interesting theorems requires a rich – and undecidable – language. In biology, useful information can be formalized in languages that can never express logical inconsistencies, e.g. in the Gene Ontology[4], and is therefore particularly suited for collaborative knowledge acquisition. A similar problem arises if the type system was reduced to a top-level ontology such as the GFO[28], in order to modify the

core ontology itself. The language in which the core ontology is formalized is at least description logic, and despite the fact that the language to formalize this knowledge would become highly complex, the problem of how to treat inconsistencies arises and ways to resolve these must be found.

Another issue to be solved before a gene function wiki could become successful is that of trust in the information in such a wiki. The advantage of the ontologies provided by a consortium is that domain and ontology experts collaborated in the creation of these ontologies which results in a high quality for them. A way to solve this in a wiki is to use a confidence or reputation system for the users of the wiki. However, the representation of this confidence in the knowledge model is problematic: it may result in fuzzy truth values for parts of the ontology[30], which would have to be defuzzicated for most applications. We also consider a web-of-trust approach[31] where users can trust particular users, e.g., ontology experts, and obtain a view on the knowledge in the wiki based on their choice of whom to trust. However, it remains open what the neutral perspective on the wiki[32] would be, as this would divide the wiki in a number of personal knowledge bases, different for each user depending on whom she trusts.

Finally, the integration of external databases and ontologies may require modifications on automated reasoners. Many ontologies which exist at present are formalized using a trivial knowledge representation language, often based on directed acyclic graphs. Reasoning on these structures is highly efficient, which is necessary as these ontologies tend to have a large number of concepts. Ontologies formalized in description logic, on the other hand, require more sophisticated reasoners which are much less efficient, while they usually have much fewer concepts. For efficiency, it would be beneficial to employ a hybrid reasoner which uses the most efficient reasoning algorithm for each part of the knowledge base. E.g., it performs a graph search on a directed acyclic graph and reuses the results from this query when performing queries which require the core ontology (which is formalized in description logic), while preserving the semantics and definitions which are given by the integration of the domain ontologies in the core ontology.

7 Conclusion

Let us revisit what we have discussed so far. First, we argued that a wiki can be enhanced by semantic relations, and that this addition is necessary for our application in order to search for genes and other biological entities, automatically check consistency, classify and group genes together, and for the integration with other knowledge bases. Second, we introduced our data model for storing the semantic content of the wiki, using n -ary relations.

Finally, a type system which is based on a formal core ontology for bio-medicine is beneficial. Because the information in the gene function wiki is highly structured due to the representation of semantic relations, it is necessary to provide the most general building blocks of the semantic content in the wiki. We will use the types provided by the core ontology as type system for the assertion of semantic relations. We can use a top-level ontology as foundation for the biomedical core ontology in order to allow the content of the gene function wiki to be used in a wider scientific context – such as chemistry.

In summary we have discussed requirements and some theoretical aspects of the implementation of a gene function wiki which we believe may provide new insights for biologists, as well as the Semantic Web and the wiki community. A work-in-progress, prototypical implementation, which, however, may not be (fully) functional, can be found for evaluation purposes on <http://onto.eva.mpg.de/bowiki>.

An implementation of a gene function wiki has the potential to provide a powerful tool for the annotation of gene data in biology. Additionally, the integration of formal ontologies and wikis may lead to new applications for wikis and ontologies in areas where their use has been rather limited until now. Further, using the framework introduced here for curation and maintenance of biomedical ontologies will enable the possibility to use information and ontology extraction methods from computer linguistics in order to create prototypical ontologies, or to generally make ontology curation and annotation faster and cost-efficient[33].

With the rapid growth of biological knowledge increasingly sophisticated methods are needed in order to close the gaps in storing, processing and representing this knowledge. Using a wiki for these purposes is far more than just an appealing idea. The needs of researchers force us to consider novel approaches such as data organization using ontologies, and data mining combined with reasoning over the given facts.

References

1. Chicurel, M.: Bioinformatics: bringing it all together. *Nature* **419**(6908) (2002) 751–755
2. Birney, E., et al.: Ensembl 2006. *Nucleic Acids Res* **34**(Database issue) (2006) 556–561
3. Wheeler, D.L., et al.: Database resources of the national center for biotechnology information. *Nucleic Acids Res* **34**(Database issue) (2006) 173–180
4. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. *Nat Genet* **25**(1) (2000) 25–29
5. Wang, K.: Gene-function wiki would let biologists pool worldwide resources. *Nature* **439**(7076) (2006) 534
6. The Internet Community: Wikipedia, the free encyclopedia. <http://wikipedia.org> (2006)
7. Giles, J.: Internet encyclopaedias go head to head. *Nature* **438** (2005) 900–901
8. Auer, S.: Powl – a web based platform for collaborative semantic web development. In: Proc. of 1st Workshop Scripting for the Semantic Web (SFSW'05). (2005)
9. Fischer, J., Gantner, Z., Rendle, S., Stritt, M., Schmidt-Thieme, L.: Semantic wiki COW. <http://www.informatik.uni-freiburg.de/cgsm/software/cow/index.en.html> (2006)
10. Völkel, M., Kröttsch, M., Vrandečić, D., Haller, H., Studer, R.: Semantic wikipedia. In: Proceedings of the 15th international conference on World Wide Web. (2006) 585–594
11. Harris, D., Harris, N.: Kendra base. <http://base.kendra.org.uk/> (2006)
12. Dello, K., Tolksdorf, R., Paslaru, E.: Makna. <http://www.apps.ag-nbi.de/makna/> (2006)
13. Campanini, S.E., Castagna, P., Tazzoli, R.: Platypus wiki: a semantic wiki web. In: Semantic Web Applications and Perspectives, Proceedings of 1st Italian Semantic Web Workshop. (2004)
14. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. *Genome Biol* **6**(2) (2005) R21

15. Burek, P., Hoehndorf, R., Loebe, F., Visagie, J., Herre, H., Kelso, J.: A top-level ontology of functions and its application in the Open Biomedical Ontologies. *Bioinformatics* **22**(14) (2006) e66–e73
16. Burek, P.: *Ontology of Functions*. PhD thesis, Institute of Informatics (IfI), University of Leipzig (2006) forthcoming.
17. Hoehndorf, R., Prüfer, K., Backhaus, M., Visagie, J., Kelso, J.: The design of a wiki-based curation system for the ontology of functions. In: *The Joint BioLINK and 9th Bio-Ontologies Meeting*. (2006)
18. Horrocks, I., Patel-Schneider, P.F.: Reducing OWL entailment to description logic satisfiability. In: *Proc. ISWC 2003*. Number 2870 in LNCS. Springer (2003) 17–29
19. Devlin, K.: *Logic and Information*. Cambridge University Press (1991)
20. Group, O.M.: UML 2.0 infrastructure specification. Document ptc/03-09-15 (2004)
21. Pepper, S., Moore, G.: XML topic maps (XTM) 1.0. <http://topicmaps.org/xtm/1.0/> (2001)
22. Garshol, L.M., Moore, G.: Topic maps – XML syntax. <http://www.isotopicmaps.org/sam/sam-xtm/> (2006)
23. Loebe, F.: *An analysis of roles: Towards ontology-based modelling*. Master's thesis, Institute of Informatics (IfI), University of Leipzig (2003)
24. Loebe, F.: Abstract vs. social roles: A refined top-level ontological analysis. In Boella, G., Odell, J., van der Torre, L., Verhagen, H., eds.: *Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems'*, Nov 3-6, Arlington, Virginia. Number FS-05-08 in *Fall Symposium Series Technical Reports*, Menlo Park (California), AAAI Press (2005) 93–100
25. Rector, A., Stevens, R., Rogers, J.: Simple bio upper ontology. <http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/> (2006)
26. Loebe, F., Hoehndorf, R.: *General Formal Ontology*. <https://savannah.nongnu.org/projects/gfo/> (2006)
27. Vrandečić, D., Krötzsch, M.: Reusing ontological background knowledge in semantic wikis. In Völkel, M., Schaffert, S., Decker, S., eds.: *Proceedings of the First Workshop on Semantic Wikis – From Wikis to Semantics*. (2006)
28. Herre, H., Heller, B., Burek, P., Hoehndorf, R., Loebe, F., Michalek, H.: *General Formal Ontology (GFO) – a foundational ontology integrating objects and processes*. *Onto-Med Report 8*, University of Leipzig (2006)
29. Masolo, C., Guizzardi, G., Vieu, L., Bottazzi, E., Ferrario, R.: Relational roles and qua-individuals. In Boella, G., Odell, J., van der Torre, L., Verhagen, H., eds.: *Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems'*, Nov 3-6, Arlington, Virginia. Number FS-05-08 in *Fall Symposium Series Technical Reports*, Menlo Park (California), AAAI Press (2005) 103–112
30. Straccia, U.: A fuzzy description logic. In: *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, Madison, US (1998)
31. Wikipedia: Web of trust — wikipedia, the free encyclopedia (2006) [Online; accessed 19-August-2006].
32. Wikipedia: Wikipedia:neutral point of view — wikipedia, the free encyclopedia (2006) [Online; accessed 19-August-2006].
33. Good, B.M., Tranfield, E., Tan, P., Shehata, M., Singhera, G., Gosselink, J., Okon, E., Wilkinson, M.: Fast, cheap and out of control: a zero curation model for ontology development. In: *Proceedings of the PSB'06*. (2006)

Using Semantic Web Tools to Integrate Experimental Measurement Data on Our Own Terms

M. Scott Marshall¹, Lennart Post^{1,2}, Marco Roos¹, and Timo M. Breit¹

¹ Integrative Bioinformatics Unit

² Nuclear Organisation Group

Institute for Informatics, Swammerdam Institute for Life Sciences

Faculty of Science

University of Amsterdam

marshall@science.uva.nl

Abstract. The -omics data revolution, galvanized by the development of the web, has resulted in large numbers of valuable public databases and repositories. Scientists wishing to employ this data for their research are faced with the question of how to approach data integration. Ad hoc solutions can result in diminished generality, interoperability, and reusability, as well as loss of data provenance. One of the promising notions that the Semantic Web brings to the life sciences is that experimental data can be described with relevant life science terms and concepts. Subsequent integration and analysis can then take advantage of those terms, exposing logic that might otherwise only be available from the interpretation of program code. In the context of a biological use case, we examine a general semantic web approach to integrating experimental measurement data with Semantic Web tools such as Protégé and Sesame. The approach to data integration that we define is based on the linking of data with OWL classes. The general pattern that we apply consists of 1) building application-specific ontologies for “myModel” 2) identifying the concepts involved in the biological hypothesis, 3) finding data instances of the concepts, 4) finding a common domain to be used for integration, and 5) integrating the data. Our experience with current tools indicates a few semantic web bottlenecks such as a general lack of ‘semantic disclosure’ from public data resources and the need for better ‘interval join’ performance from RDF query engines.

1 Introduction

The -omics data revolution, galvanized by the development of the web, has produced large numbers of valuable public databases and repositories. These databases enable many types of research by providing free web access to essential up-to-date -omics information and even raw data. However, the same revolution has also led to an explosion of proprietary formats and interfaces. Researchers who want to integrate data from several sources must find a way to extract information from a variety of search interfaces, web page formats, and API's. To complicate matters, some databases periodically change their export formats, effectively breaking the tools that provide access to their data. Although this scenario is an improvement on a decade

ago, there is still very little Semantic Web technology involved. Most -omics databases do not yet provide metadata and, when it is available, do not provide it in a standard format with common semantics. We envision a future where not only data but also schemas that describe them are accessible in semantic web formats such as RDF, RDFS, and OWL, and data is provided with the semantic annotations that are necessary to link them to the concepts that describe their components.

One of the most promising notions that Semantic Web brings to biology is that the search for data, and even experiments themselves, can eventually be specified in terms of the relevant biological concepts. Once disclosed along with the corresponding data, these concepts can then serve as part of the documentation for the experiment itself, helping to encode the hypothesis and relevant domain knowledge. Moreover, these concepts can eventually be used to define and steer the execution of a computational experiment, thus removing the burden of many implementation details and allowing scientists to define experiments in their own terms, i.e. ontologies of their choice or making.

An essential element of a semantic web contribution for the life sciences is data integration. Most forms of computational biology, workflow, data analysis, and visualization require data integration (see [1, 2] and references therein). In the context of the Virtual Laboratory e-science (VL-e) project, we consider how biologists could perform integrative bioinformatics research by considering biology use cases as working examples. Our specific use case requires data integration in order to explore the viability of a hypothesis that links epigenetics and transcription. Our goal is to perform data integration in a way that is repeatable and self-documenting as a result of syntactic and semantic disclosure. In this article, we describe a semantic web approach to performing biological data integration that we think is general enough to be useful to a variety of disciplines in the context of virtual laboratories and e-science.

1.1 Biology Use Case-Background

The goal of our use case is to unravel the relationship between the histone code, DNA sequence, and transcription. We start our incremental approach by studying the relationship between two components: histones and transcription factor binding sites in the DNA sequence. Histones are specific types of proteins that bind DNA and as such are central to packaging long DNA molecules into chromosomes in the nucleus of a cell. They undergo specific modifications, such that a pattern over the chromosomes is formed, referred to as a ‘histone code’ [3]. ‘Transcription factors’ are also proteins that can bind DNA to directly influence gene expression. Many of the DNA sequences to which transcription factors bind, i.e. transcription factor binding sites, have been identified and localized on human DNA. The biological question in our case is: How is chromatin involved with transcription?

1.2 Creating Application-Specific Ontologies for *myModel*

The first step of our approach is to assemble the concepts relevant to our biological hypothesis. These concepts will serve as the terms of a controlled vocabulary that we

can use to build our queries. The use of ontologies as controlled vocabularies can be found in practice such as in [4]. Lacking an existing ontology that covers the concepts relevant to our research problem, we create an ontology that will serve as *myModel*¹. This ontology is limited to the concepts necessary to describe the problem domain of our experiment and could be called an *application-specific ontology*. We expect our ontology to evolve during the course of our case studies, or be merged with ontologies created by the domain community. These same OWL classes can be eventually used to link to other knowledge, such as relations with other OWL classes or rules [5] for use by reasoners. Although the ontology is used as a namespace during our data integration, the consistency checking that is possible is an important advantage of using OWL for *myModel*. Of course, best practices and design patterns [6] should be employed to ensure correctness and reusability. To enhance future interoperability with biomedical ontologies, we are investigating how to employ the relations proposed by Smith et al [7].

After evaluating the Gene Ontology [8] and seeking other appropriate ontologies using Swoogle [9], we decided to build our own application-specific ontology for histones in OWL: *HistOn*. One of the main reasons for doing this was to include a level of detail related to histones that we did not encounter in existing ontologies. We used the OWL plug-in of Protégé [10, 11]. To facilitate future reuse of the major parts of *HistOn* we created separate OWL files for each (the combined ontology can be viewed at [12]):

myModel consists of the following ontologies:

- Higher-level concepts related to epigenetics², such as *ChromosomeRegion*
- Histones and concepts directly related to histones
- Transcription factor binding sites and directly related concepts

1.3 Asserting the Hypothesis

To represent the hypothesis for our use case, we started by drawing a cartoon, in line with common practice in the field of biology ([13]; Figure 1a). It shows, in ‘biologist-readable’ form, that we want to study the relationship between a particular histone (H3) with a particular modification (tri-methylation on the fourth lysine, i.e. H3K4Me3), and transcription factor binding sites (TFBS) because they are related to gene expression, and that both these elements are located on the DNA of a chromosome. In contrast to common practice in biology, we added concepts from the cartoon to *myModel*. Once we have added these concepts to *myModel* we can define the hypothesis in terms of this model (Figure 1b). In our example, the hypothesis is simply that there is a relation between the two concepts of chromatin and transcription. Knowledge representation such as that for hypothetical assertions remains future work.

¹ We use “*model*” to mean the machine-readable qualitative model relevant to the phenomenon being studied.

² Epigenetics refers to the heritable control over gene expression that is not linked to the DNA sequence alone; histones are likely to play an important role in this control.

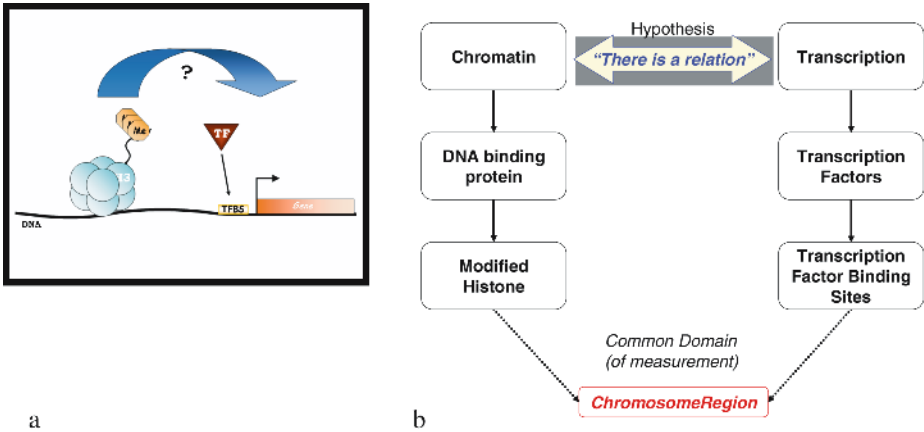


Fig. 1. (a) Cartoon representation (b) Schematic overview of the path to a common domain

1.4 Finding Relevant Data

Lacking a data broker or semantic mediator (see *Related Work*), most biologists must rely on their incidental knowledge of web resources to find appropriate data sources. In our case, this led us to the UCSC genome browser [14, 15]. The curators use an annotation strategy based on the genomic alignment algorithm BLAT [16]. In principle, any biological entity that can be associated with a DNA sequence can be localized on the chromosomes (human, rat, and mouse are among the species available at UCSC). Data from the ENCODE project, including histone binding data, is stored here, as well as transcription factor binding sites (TFBS).

1.5 Data Import

We would like our final RDF data to contain information about original table structure, data source, the entry or row, syntactic type, and semantic type. We separate the import process into two steps: a syntactic and a semantic annotation step.

The syntactic step makes use of table information provided by UCSC. We defined an OWL schema (*theirDataModel*) to represent the table structures, both for the histone data and the transcription factor binding sites. This part of the import is similar to that used in YeastHub [17], where table column names are translated directly into RDF types, resulting in an RDF version of the UCSC table structure for the data. Each row of the table corresponds to a single measurement and requires a unique identifier, which we generate based on source file and row number. In this particular use case, we are not confronted with the more difficult but related problem of using a globally unique identifier (e.g. for a gene) such as is being discussed in the research community (see for example related discussions in [18]) . Our measurement identifier is only required to be unique over the set of measurements that are being integrated. We also used the type information provided by UCSC in the form of a MySQL dump to generate the appropriate XML Schema Definition (XSD) tag for

each piece of data³. The XSD tag was to ensure that our data was properly interpreted (for example, not as a string in a comparison meant for numbers). The translation of the UCSC data into RDF was performed using a version of Mapper [19] that we modified for RDF output [20].

The semantic annotation step of the import process requires a mapping from the RDF types in the data (*theirDataModel*) to the corresponding semantic types in *myModel* (a model from the data producers, *theirModel*, has not been supplied). The purpose of the mapping is to enable our “semantic query” to be made in terms of *myModel* so we want to create the conditions where our query selects data with *theirDataModel* types when the query actually contains our own *myModel* types. We found that a subclassing of a *theirDataModel* property to a *myModel* property with `rdfs:subPropertyOf` produced the desired effect, both for that property and the RDF nodes at its endpoints (due to the RDFS reasoning in Sesame). This type of subtype mapping should also be possible in the case that a data provider supplies semantic metadata (i.e. *theirModel*) although the required ‘ontology alignment’ could be more complex⁴. Note that we could have directly translated column headers into the corresponding equivalent OWL type during the Mapper to RDF step and make RDFS reasoning unnecessary. However, such an approach would shift control of the mapping from the RDFS statements to the Mapper import stage and subsequent changes to *myModel* could require building an entire new RDF graph of the data with the new names that have resulted from the changes.

1.6 A Basis for Comparison: Finding the Common Domain for Integration

In order to integrate measurement data, we must align values along the same domain or axis. We will take our use case as an example. We can use the graph of our ontology to look for such a domain. The comparable domain in our use case is the region defined by the class ‘*ChromosomeRegion*’. A chromosome region is an interval of DNA sequence located along a particular chromosome. In terms of the concept graph, *ChromosomeRegion* forms the link between the two concepts that we want to compare: both histones and transcription factor binding sites are related to this concept.

We also need to establish the criteria that make a given pair of measurements comparable, i.e. the measurements should be sampled from the same part of the domain. When there is overlap in the measurement domain, we want to compare the measurement values from the two different data sources corresponding to the overlap. To begin with, we chose a simple overlap criterion for our *ChromosomeRegion* intervals that we could encode directly in an RDF query.

1.7 Data Integration Query

Once we have found a way to determine which measurement data can be meaningfully paired, we can perform the final step of our data integration experiment. Although it is possible to write a program to achieve this step, we chose to write an RDF query (see [21] for details). In this way, the semantics of the integration are

³ We ran into a technical problem for large numbers (> 6M) of XSD tags and describe the solution at <http://integrativebioinformatics.nl/histone/HistoneDataIntegration.html>.

⁴ Ontology alignment is an area of ongoing research.

readily available for inspection: all terms used in the query refer to OWL classes. With RDFS reasoning on, we can take advantage of the subsumption equivalence to our own myModel names, i.e. write the query with our own OWL “terms”. Our query returns a list of data pairs that can then be further explored by browsing, visualization, statistics (e.g. correlation), and data mining. Our preliminary results suggest that a large number of TFBS types are preferentially located within the regions overlapping H3K4Me3 binding sites, which is in line with experiments that suggest a role for H3K4Me3 in gene activation [22]. Further biological characterization of these TFBSs is work in progress.

1.8 Performance Issue for ‘Interval Join’

The type of query that we use is called an ‘interval join’ in temporal and multimedia databases, where regions of media are checked for overlap with the regions defined for corresponding annotations. The ‘interval join’ appears to be unavoidable when performing data integration of measurements by query. Our largest datafile (all data for the genome) contains approximately 11M triples. Scalability and performance issues arose during initial tests, forcing us to run with smaller data and try different configurations. In this phase, we did not use RDFS reasoning and started performing queries in terms of *theirDataModel*. We created test data from a smaller data set (chromosome 22), and tried the query in several RDF systems⁵. It has been suggested [23] that different combinations of data and query can produce widely varying results. We indeed found that our query/data scaled unpredictably depending on the RDF implementation being used, with our query on the full data set taking on the order of days (see Table 1 and *Disclaimer*). In contrast, several non-RDF implementations executed the query in a matter of seconds. This discrepancy and significant performance differences between the RDF implementations themselves points to a performance bottleneck that could be better supported in RDF query engines, perhaps with custom support for our particular type of join. Note that in the case of MonetDB, custom optimizations for interval joins (called “StandOff joins” in [24]) can result in dramatic

Table 1. Naïve comparison (unequal platforms and technologies)

	Chromosome 22	Genome
SWI Prolog (RDF)	3.25 minutes	X
Sesame (RDF)	4.25 minutes	42 hours
Jena (RDF)	8 minutes	8 days
Python custom program	X	17 seconds
XQuery (MonetDB)	X	7 seconds
mySQL	X	98 minutes

Disclaimer: This table is a naïve comparison and not a benchmark! Different conditions exist between tests e.g. machines, machine load, level of configuration and API expertise, version numbers, etc.

⁵ Although the terms of our license agreement do not allow us to publish the performance results in our table at this time, our tests with the RDF implementation of an anonymous major DB vendor produced no improvements.

speedup in XQuery. However, although MySQL apparently performs better than RDF implementations with our query, the ‘interval join’ would still be too demanding for a public server based on a MySQL database⁶. Although space limitations prevent us from including the text of an actual query here, example queries can be found at [21].

2 Related Work

Data integration is an important topic in biology, and numerous solutions have been developed to enable retrieval of data from heterogeneous distributed sources (for reviews see e.g. [1, 2, 25]). The solutions range from monolithic, such as SRS [26] that uses keyword indexing and hyperlinks, Kleisli/K2 [27] that uses a query language that can query across databases as if they were one, data warehouses such as BioZon [28], to solutions that use web services acting as portals to biological data [29]. Perhaps the most widely used system is SRS, providing integration of more than 400 databases.

Our approach to data integration uses semantic models to provide a schema for integration. TAMBIS pioneered such an approach by creating a molecular biology ontology as a global schema for transparent access to a number of sources including Swiss-Prot and Blast [30]. Systems such as BACIIS [31], BioMediator [32] and INDUS [33] extend on this example. For instance, BioMediator uses a ‘source knowledge base’ that represents a ‘semantic web’ of sources linked by typed objects. The knowledge base includes a ‘mediated schema’ that can represent a user’s domain of discourse. INDUS shows important similarities to our approach. INDUS offers an integrated user interface to import or create user-ontologies (similar to ‘myModel’, but limited to ‘attribute-value hierarchies’), and create ontological mappings between concepts and ‘ontology-extended’ data sources. In contrast to our approach, however, INDUS does not use semantic web formats such as OWL and RDF. While the syntactic step of our import is similar to that of YeastHub [17], our explicit linking of the semantic types to the syntactic types with RDFS moves the work of discovering the semantics from the query stage to the stage of model alignment.

3 Future Work

Our import process is meant to eventually create a transparent data access layer (termed *wrapper* in BACIIS) to external data sources such as UCSC. The import approach that we have described here is being extended to work on all UCSC data. The translation to RDF can be automated with the use of the MySQL table information provided by UCSC. The table information can be used to create both the *theirDataModel* in OWL and the XML “map” used by the Mapper program for a standardized mapping from UCSC data to RDF. Of course, our approach is general and can eventually be applied to other data providers. However, we expect that data providers such as UCSC will add RDF export to their set of services. Once RDF export is available, a general data mediation (web) service becomes possible.

⁶ This could be done with the translation from RDF in the case of a query rewriting approach such as that employed for D2RQ [34].

Approaches that make use of a mapping between RDF and relational database schemas such as D2RQ [34] could eventually be used to provide data access via RDF queries.

Our query results in a set of abstract *overlaps* or derived features, i.e. regions of a domain from which measurements have been taken that we have deemed ‘interesting’ according to a criterion (in our case, overlap in the domain of measurement). These *overlaps* contain information relevant to our hypothesis. A challenge for computational experimentation is to create an OWL class for use in the semantic annotation of these *overlaps* that exposes them to queries for data related to *Chromatin* and *Transcription*.

4 Discussion

We propose a semantic web approach to data integration and report on our experience applying it in the context of a biological use case. Our approach is model-oriented and allows us to perform data integration experiments in terms of our own biological knowledge. It allows us to perform data integration of experimental measurement with a query in terms of *myModel*. This type of ‘semantic disclosure’ exposes meaning and application logic that would otherwise only be available to scientists that can interpret the code of the application that uses the data.

It appears that an interval join is unavoidable wherever measurement data is to be integrated with a query language. Better support for interval joins in RDF query engines are therefore important for adoption of this approach for exploratory analysis, where interactivity is generally preferred.

One of the semantic web bottlenecks that we have encountered is the general lack of semantic disclosure: i.e. *theirModel* (semantic model) is not supplied by the data provider. Such information is especially crucial to efforts that attempt to facilitate data integration that crosses domains of expertise. A more practical reason is that in some cases it is difficult to discover what the biological data really means. Therefore, we find it encouraging that *theirModel* could become available from, for example, NCBI [35] in the future.

Acknowledgements

We thank Willem van Hage for his assistance with RDFS features of Sesame. Thanks to Peter Boncz, Bart Heupers, Jacco van Ossenbruggen, and Jan Wielemaker (as well as colleagues at the anonymous major DB vendor) for help with the tests in Table 1.

This work was carried out in the context of the Virtual Laboratory for e-Science project (<http://www.vl-e.nl>), and the BioRange program of the Netherlands Bioinformatics Centre (NBIC). VL-e is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ). BioRange is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

References

1. Searls DB. Data integration: challenges for drug discovery. *Nat Rev Drug Discov* 2005; 4(1): 45-58.
2. Stein LD. Integrating biological databases. *Nat Rev Genet* 2003; 4(5): 337-45.
3. Strahl BD and Allis CD. The language of covalent histone modifications. *Nature* 2000; 403(6765): 41-5.
4. **About BIRNLex** [<http://xwiki.nbirn.net:8080/xwiki/bin/view/BIRN-OTF/About+BIRNLex>]
5. **Rule Interchange Format Working Group Charter** [<http://www.w3.org/2005/rules/wg/charter>]
6. **SWBP&D WG Semantic Web Tutorials** [<http://www.w3.org/2001/sw/BestPractices/Tutorials>]
7. Smith B, et al. Relations in biomedical ontologies. *Genome Biol* 2005; 6(5): R46.
8. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25(1): 25-9.
9. Ding L, et al. Swoogle: a search and metadata engine for the semantic web. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM Press: Washington, D.C., USA. 2004. 652-659.
10. **Protégé** [<http://protege.stanford.edu/>]
11. Knublauch H, Dameron O, and Musen MA. Weaving the Biomedical Semantic Web with the Protégé OWL Plugin. *First International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004)* 2004 (Whistler (BC, Canada)), American Medical Informatics Association; 33-47.
12. **OWLDocs of Overview Ontology for myModel** [<http://integrativebioinformatics.nl/histone/OWLDocs/OverviewOntology/index.html>]
13. Perini L. Explanation in Two Dimensions: Diagrams and Biological Explanation. *Biology and Philosophy* 2005; 20: 257-269.
14. Gribskov M. Challenges in data management for functional genomics. *Omics* 2003; 7(1): 3-5.
15. Kent WJ, et al. The human genome browser at UCSC. *Genome Res* 2002; 12(6): 996-1006.
16. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res* 2002; 12(4): 656-64.
17. Cheung KH, et al. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* 2005; 21 Suppl 1: i85-i96.
18. **Semantic Web for the life sciences discussion forum** [<http://lists.w3.org/Archives/Public/public-semweb-lifesci/>]
19. **Navigate data with the Mapper framework, Build your own data mapping system with an interlingual approach** [<http://www.javaworld.com/javaworld/jw-04-2002/jw-0426-mapper.html>]
20. **Mapper** [<https://gforge.vl-e.nl/projects/mapper/>]
21. **Semantic Data Integration for Histone Use Case Website** [<http://integrativebioinformatics.nl/semanticdataintegration.html>]
22. Schubeler D, et al. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev* 2004; 18(11): 1263-71.
23. **Pitfalls in Benchmarking Triple Stores** [<http://jeenbroekstra.blogspot.com/2006/02/pitfalls-in-benchmarking-triple-stores.html>]
24. Alink W, et al. Efficient XQuery Support for Stand-Off Annotation. *Proceedings of International Workshop on XQuery Implementation, Experience and Perspectives (XIME-P)* 2006 (Chicago, IL, USA).

25. Eckman B, Rice J, and Schwarz P. Data management in molecular and cell biology: vision and recommendations. *Omics* 2003; 7(1): 93-7.
26. Zdobnov EM, et al. The EBI SRS server-new features. *Bioinformatics* 2002; 18(8): 1149-50.
27. Ritter O, et al. Prototype implementation of the integrated genomic database. *Comput Biomed Res* 1994; 27(2): 97-115.
28. Birkland A and Yona G. BIOZON: a hub of heterogeneous biological data. *Nucleic Acids Res* 2006; 34(Database issue): D235-42.
29. Wilkinson M, et al. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol* 2005; 138(1): 5-17.
30. Stevens RD, Robinson AJ, and Goble CA. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* 2003; 19 Suppl 1: i302-4.
31. Ben Miled Z, et al. An efficient implementation of a drug candidate database. *J Chem Inf Comput Sci* 2003; 43(1): 25-35.
32. Mork P, Shaker R, and Tarczy-Hornoch P. The Multiple Roles of Ontologies in the BioMediator Data Integration System. *DILS* 2005, Springer; 96-104.
33. Caragea D, et al. Algorithms and Software for Collaborative Discovery from Autonomous, Semantically Heterogeneous, Distributed Information Sources. *ALT* 2005, Springer; 13-44.
34. **D2RQ** [<http://www.wiwiss.fu-berlin.de/suhl/bizer/d2rq/spec/>]
35. **public-semweb-lifesci forum message from Benjamin H. Szekely** [<http://www.w3.org/mid/OFC5D7E901.5F3825EB-ON85257169.0060CA27-85257169.006B0FEE@us.ibm.com>]

Ontology Guided Data Integration for Computational Prioritization of Disease Genes*

Bert Coessens¹, Stijn Christiaens², Ruben Verlinden², Yves Moreau¹,
Robert Meersman², and Bart De Moor¹

¹ Department of Electrical Engineering
Katholieke Universiteit Leuven

`bert.coessens, yves.moreau, bart.demoor@esat.kuleuven.be`

² Semantics Technology and Applications Research Laboratory
Vrije Universiteit Brussel

`stijn.christiaens, ruben.verlinden, robert.meersman@vub.ac.be`

Abstract. In this paper we present our progress on a framework for collection and presentation of biomedical information through ontology-based mediation. The framework is built on top of a methodology for computational prioritization of candidate disease genes, called Endeavour. Endeavour prioritizes genes based on their similarity with a set of training genes while using a wide variety of information sources. However, collecting information from different sources is a difficult process and can lead to non-flexible solutions. In this paper we describe an ontology-based mediation framework for efficient retrieval, integration, and visualization of the information sources Endeavour uses. The described framework allows to (1) integrate the information sources on a conceptual level, (2) provide transparency to the user, (3) eliminate ambiguity and (4) increase efficiency in information display.

1 Introduction

The ever increasing amount of biological data and knowledge, its heterogeneous nature, and its dissemination all over the Internet, make efficient data retrieval

* Research supported by: [Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymbioSys, IDO (Genetic networks), several PhD/postdoc & fellow grants]; [Flemish Government: FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (Bonatema), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel) research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos]; [Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modeling, 2002-2006)]; [EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain].

a horrendous task. Biological research has to deal with the diversity and distribution of the information it works with [1]. Yet, access to and integration of a multitude of complementary data sources will become critical to achieve more global views in biology.

Current solutions to these problems are usually handled manually and integration becomes a tedious activity of ad-hoc implementation [2]. Part of these problems originate from the fact that most of these data sources are created as if they exist alone in the world [3]. It is clear that a good framework is a necessity, in which integration can be done through configuration, rather than case-specific implementation. Such a framework will need to be based on semantics, rather than on syntax and structure [4].

Integration on a conceptual level also opens up new opportunities for creating information-rich user interfaces. Presenting a user with the information she needs, augmented with relations to relevant data, is an approach used commonly in semantic web browsers (e.g., *OntoWeb* [5]) and even in industry software (e.g., *Context Browser* [6]). *Dzbor*, *Domingue*, and *Motta* [7] call this an interpretative viewpoint or also context. This kind of interface allows a disambiguating stepwise refinement of the user's search [8].

In contrast to warehouse oriented tight integration systems (e.g., *Biozon* [9]), our approach is to provide efficient navigation through sets of loosely integrated data sources. The purpose of our system is not to allow generic queries on overlapping data sources, but rather to give transparent access to information that is highly relevant and useful for the analysis at hand. Thus, the system fills a very specific need in relieving the researcher of the burden to manually collect or navigate to all necessary information.

According to the classification proposed by *Hernandez* and *Kambhampati* [1], our system falls in the category of portal systems (like *SRS* [10]), but has a structured object-relational data model (like *TAMBIS* [11], *K2/BioKleisli* [12], *DiscoveryLink* [13], etc.). The system is limited to horizontal integration of data sources, but being navigational it does not require critical expertise of a specific query language (like in systems using *CPL* [14,11] or *OQL* [15,12], for instance).

We start with a description of *Endeavour* and the problem of computational gene prioritization. In section 3 we give a brief overview of the *DOGMA* framework for ontology engineering, followed by an explanation how an ontology can be used for mediation and visualization of the information used in the prioritization methodology in section 4. We show how this solves several integration problems. We end this paper with some conclusions and possible future directions in section 5.

2 Endeavour

In the field of linkage analysis and association studies researchers are often confronted with large lists of candidate disease genes, especially when investigating complex multigenic diseases. Investigating all possible candidate genes is a tedious and expensive task that can be alleviated by selecting for analysis only the

most salient genes. Also, in the context of high-throughput experiments (like microarray gene expression assays), ever growing amounts of gene-associated data make manual investigation of *interesting* genes nearly unfeasible. It is clear that efficient and statistically sound computational prioritization methods become increasingly important.

With respect to this need, we developed the Endeavour methodology for the computational prioritization of a group of candidate genes based on their similarity with a set of training genes [16]. The method uses Order Statistics to combine a variety of information sources and clearly has several advantages over other approaches. It solves the problem of missing data and reconciles even contradictory information sources. It allows for a statistical significance level to be set after multiple testing correction, thus removing any bias otherwise introduced by the expert during manual prioritization. It also removes part of the bias towards known genes by including data sources that are equally valid for known and unknown genes. The methodology has been validated in a large scale leave-one-out experiment with 29 diseases and 627 disease genes fetched from OMIM. On top of that, several prioritizations were validated in wet-lab experiments. These analyses were published in Nature Biotechnology by Aerts *et al.* [16].

Endeavour Prioritization Terminology. The central object in the Endeavour prioritization methodology is a *Gene*. This object represents a biological entity and all information known about it. In most of the cases, this entity will be a gene. Biological entities are combined in sets (*GeneGroup*). A training set is a *GeneGroup* that is used to build a model for a process or disease, represented by the *Model* object. A *Model* consists of several *SubModel* objects that each represent a certain data source. Building a *SubModel* means fetching and summarizing all information about the genes in the training set for one particular data source.

Endeavour comes with a set of standard submodels that summarize the following information about the user-specified training genes: KEGG pathway membership [17], Gene Ontology (GO) annotations [18], textual descriptions from MEDLINE abstracts, microarray gene expression, EST-based anatomical expression, InterPro's protein domain annotation [19], BIND protein interaction data [20], *cis*-regulatory elements, and BLAST sequence similarity. Besides these default information models, users can add their own microarray data or custom prioritizations as submodels. Most of the data sources are either vector-based (e.g., textual information, gene expression data) or attribute-based (GO, EST, InterPro, Kegg).

Apart from the *GeneGroup* that contains the training genes, there is a second *GeneGroup* that holds the candidate genes and all their related information. These candidate genes are prioritized during a process called *scoring*. Scoring involves comparing the information of a candidate gene with the information in the *Model* object for every data source. Based on these comparisons, every candidate gene receives a ranking. All rankings of the test genes according to the different available data sources are then combined using order statistics to obtain one overall ranking.

Endeavour Information Browser. The decision was taken to provide the users of Endeavour with a maximal control over the set of training and test genes, as well as over the data sources to include in the prioritization. This idea was conceived from a prospective discussion with many geneticists and biologists, who do not use the existing prioritization methods for their lack of flexibility. This is perhaps best illustrated by the fact that not a single paper has been published reporting the identification of a novel disease gene when using any of the pre-existing methods. Most likely, this relates to the reality that geneticists and biologists, as opposed to bioinformaticians, prefer to have the flexibility to interactively select their own set of genes and the information they want to work with, above an automatic and non-interactive data mining selection procedure of disease characteristics.

In this context, it is of utmost importance to make well-informed decisions about which genes and information sources to include in the prioritization. A user must be able to browse the relevant information efficiently and in accordance with the methodology's demands. To live up to this need, and given the heterogeneous nature of the biological information to be consulted, an information browser was developed based on the Endeavour methodology. The existing data model was extended to a full-fledged ontology with *Gene* as the central object to allow ontology-guided browsing through the available information (see Figure 2).

3 DOGMA Ontology Paradigm

DOGMA¹ is a research initiative of VUB STARLab where various theories, methods, and tools for building and using ontologies are studied and developed. A DOGMA inspired ontology is based on the classical model-theoretic perspective [21] and decomposes an ontology into a lexon base and a layer of ontological commitments [22,23]. This is called the principle of double articulation [24].

A lexon base holds (multiple) intuitive conceptualization(s) of a particular domain. Each conceptualization is simplified to a *representation-less* set of context-specific binary fact types called lexons. A lexon represents a plausible binary fact-type and is formally described as a 5-tuple $\langle V, \text{term1}, \text{role}, \text{co-role}, \text{term2} \rangle$, where V is an abstract context identifier, lexically described by a string in some natural language, and is used to group lexons that are logically related to each other in the conceptualization of the domain. Intuitively, a lexon may be read as: within the context V , the term1 (also denoted as the header term) may have a relation with term2 (also denoted as the tail term) in which it plays a role, and conversely, in which term2 plays a corresponding co-role. Each (context, term)-pair then lexically identifies a unique concept. A lexon base can hence be described as a set of plausible elementary fact types that are considered as being true. Any specific (application-dependent) interpretation is moved to a separate layer, i.e., the commitment layer.

¹ Developing Ontology-Grounded Methods for Applications.

The commitment layer mediates between the lexon base and its applications. Each such ontological commitment defines a partial semantic account of an intended conceptualization [25]. It consists of a finite set of axioms that specify which lexons of the lexon base are interpreted and how they are visible in the committing application, and (domain) rules that semantically constrain this interpretation. Experience shows that it is much harder to reach an agreement on domain rules than one on conceptualization [28]. For instance, the rule stating that each gene is identified by an Ensembl Gene ID [26] may hold in the Universe of Discourse (UoD) of some application, but may be too strong in the UoD of another application (e.g., Entrez Gene [27]). A full formalization of DOGMA can be found in De Leenheer, Meersman, and de Moor [29,30].

4 Ontology Guided Mediation

As we are trying to integrate several heterogeneous data sources used in the Endeavour prioritization process and with conceptually overlapping instances, the approach described by Verheyden and Deray [31,32,33] is suited for our purposes. This approach uses ontologies as a mediating instrument for conceptual data integration.

Each data source is individually mapped (or committed) to the ontology using the Ω -RIDL commitment language [32]. These individual mappings ($\Omega_1 \dots \Omega_n$ in Figure 1) enable the mediator to access and query the respective wrappers according to its own view on the data. At this point, the data sources were mapped to the ontology manually.

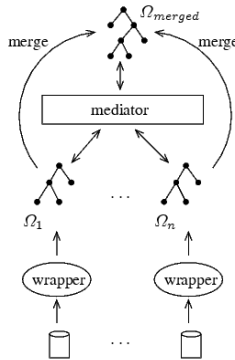


Fig. 1. Mediator Approach for Data Integration

As a proof of concept, we used this approach to integrate different, but partially overlapping, data sources from Endeavour. In close cooperation with a domain expert, we modeled the ontology based on two major types of data sources in more detail (vector- and attribute-based sources). Since the ontology is aligned directly to the Endeavour terminology, all concepts are unambiguously defined.

The obtained ontology is displayed in Figure 2 in the form of a NORM-tree [34]. Terms (e.g., gene) can be observed multiple times (in darker grey) in this representation. Each node in the tree is extended (using a double click-action) with all its related terms in order to display a complete local context at all times. When a node is selected, it forms a path (darker line in Figure 2) to the root of the NORM-tree. This local context approach is one of the abstraction mechanisms identified by Halpin [35].

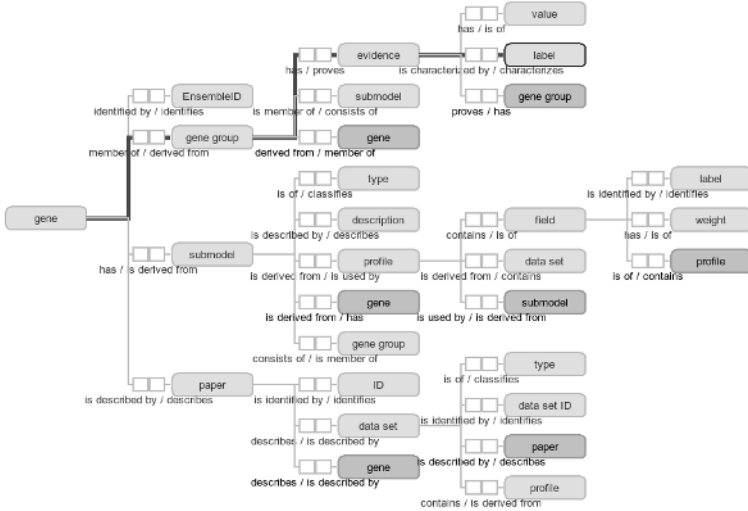


Fig. 2. Partial Endeavour Ontology visualized using the DOGMA NORM-tree representation

Figure 3 shows how the link between the different heterogeneous data sources to the conceptual model can be used to provide transparency, eliminate ambiguity, and increase efficiency when displaying relevant information to the user. Using the view in the upper part of the screenshot the user can browse the NORM-tree as described for Figure 2. The gene-related information is shown to the user in the lower half of the screen. While the user browses the NORM-tree, she selects relevant ontological paths. This way the active instance of query results provide the information she requests.

As a result of using this integrated interface the user is not confronted with the terminology and jargon used in the different data sources. The user only sees well-known terminology in the ontology when selecting relevant objects, thus eliminating ambiguity. She is also not confronted with the specific source of the data unless explicitly desired, thus providing transparency. By selecting ontological paths in the NORM-tree only information relevant to the local context is shown, thus enhancing efficiency. Eliminating ambiguity, providing transparency and enhancing efficiency to the user when browsing information relevant to a group of genes will allow her to concentrate completely on the prioritization analysis.

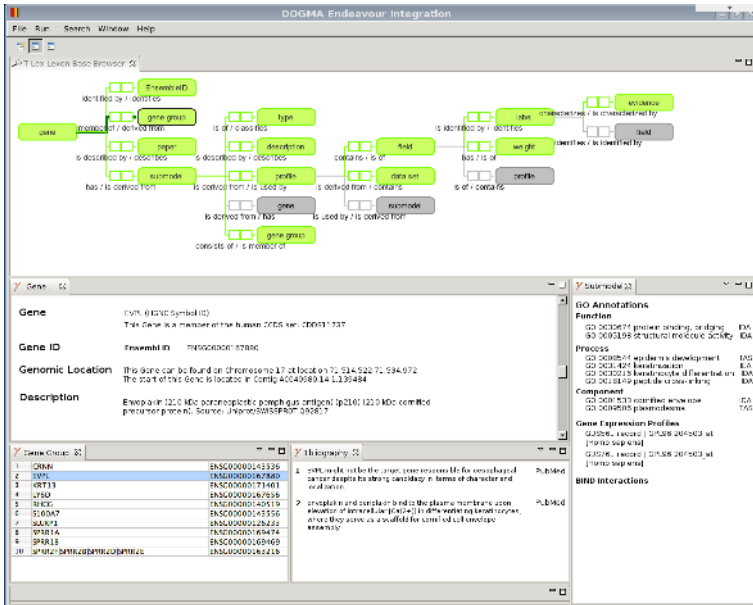


Fig. 3. Screenshot of the Endeavour Information Browser. In the top pane, the ontological paths are selected using the DOGMA NORM-tree representation. Relevant data, according to the selected paths, is retrieved on the fly from the different data sources and displayed in the lower panes. The concepts in the ontology align directly to the Endeavour terminology, which results in more efficient browsing through all information related to the analysis.

5 Discussion and Future Work

By using a conceptual approach for data integration we obtain several significant advantages. Although tedious, the mediating process is relatively easy to apply on new vector- and attribute-based data sources. The only difficulty arises if the data source (of a yet unmet type) contains data that cannot be mapped to the ontology. In this case, the ontology must be extended in order to integrate the new data. Semi-automated mapping of new data sources lies beyond the scope of the current research.

Another benefit of our approach is found in the visualization of the data. Since all data is linked (mapped) to a certain concept in the ontology, it is possible to enrich the view (e.g., the result of a query) with relevant and meaningful related information. The data is not only presented, it is also displayed in its own local context. This results in a complete transparency to the user and a more efficient visualization, as what needs to be seen, is shown.

¹ The screenshot in Figure 3 is a partial mock-up. The actual link between the gene related information and the interface still needs to be implemented.

The ontology we use can also be used by other applications, who use Endeavour itself as the data source. The conceptual model solves interoperability problems, as from the model alone, it is clear what Endeavour can provide, and how it can provide this data.

The tool is at this point a supporting application for the Endeavour system. We will extend it in order to have it actually send data to Endeavour to further facilitate gene prioritization.

References

1. Thomas Hernandez and Subbarao Kambhampati : Integration of biological sources : current systems and challenges ahead. In *ACM Sigmod Record*, 2004, 33(3), pp. 51-60
2. L. Stein : Creating a bioinformatics nation. In *Nature* 6885, 2002, 417, pp. 119-120
3. M. Stonebraker : Integrating Islands of Information , *EAI Journal*, Sept 1999. <http://www.eaijournal.com/DataIntegration/IntegrateIsland.asp>.
4. Sheth A. 1998 : Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, in *Interoperating Geographic Information Systems*, M.F. Goodchild, M.J. Egenhofer, R. Fegeas, and C.A. Kottman (eds) Kluwer Publishers
5. P. Spyns, D. Oberle, R. Volz, J. Zheng, M. Jarrar, Y. Sure, R. Studer, R. Meersman : *OntoWeb - a Semantic Web Community Portal*. In *Proc. Fourth International Conference on Practical Aspects of Knowledge Management (PAKM)*, December 2002, Vienna, Austria, 2002.
6. Context Browser : http://www.agilense.com/p_artifactmgt.html
7. M. Dzbor, J. Domingue and E. Motta : Magpie - Towards a Semantic Web Browser. In *International Semantic Web Conference 2003*, LNCS 2870, pp. 690-705
8. E. Garca and M.A. Sicilia : User Interface Tactics in Ontology-Based Information Seeking. In *Psychology e-journal* 2003 1(3):243-256.
9. Birkland, A and Yona, G : BIOZON: a system for unification, management and analysis of heterogeneous biological data. In *BMC Bioinformatics*, 2006, 7, pp. 70-70
10. T. Etzold and A. Ulyanov and P. Argos : SRS: information retrieval system for molecular biology data banks. In *Methods Enzymol*, 1996, 266, pp. 114-128
11. R. Stevens and P. Baker and S. Bechhofer and G. Ng and A. Jacoby and N.W. Paton and C.A. Goble and A. Brass : TAMBIS: transparent access to multiple bioinformatics information sources. In *Bioinformatics*, 2000, 16, pp. 184-185
12. S.B. Davidson and J. Crabtree and B.P. Brunk and J. Schug and V. Tannen and G.C. Overton and C.J. Stoeckert : K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. In *IBM Systems Journal*, 2001, 40, pp. 512-531
13. L.M. Haas and P.M. Schwarz and P. Kodali and E. Kotlar and J.E. Rice and W.C. Swope : *DiscoveryLink: A system for integrated access to life sciences data sources*. In *IBM Systems Journal*, 2001, 40, pp. 489-511
14. L. Wong : *The Collection Programming Language - Reference Manual*. Technical Report Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613
15. A. M. Alashqur and Stanley Y. W. Su and Herman Lam : OQL: A Query Language for Manipulating Object-oriented Databases In *Proceedings of the Fifteenth International Conference on Very Large Data Bases*, August 22-25, 1989, Amsterdam, The Netherlands, pp. 433-442

16. S. Aerts and D. Lambrechts and S. Maity and P. Van Loo and B. Coessens and F. De Smet and L.C. Tranchevent and B. De Moor and P. Marynen and B. Hassan and P. Carmeliet and Y. Moreau : Gene prioritization via genomic data fusion. In *Nat Biotechnol*, 2006, 24, pp. 537-544
17. M. Kanehisa and S. Goto and S. Kawashima and Y. Okuno and M. Hattori : The KEGG resources for deciphering the genome. In *Nucleic Acids Res.*, 2004, 32, pp. D277-D280
18. The Gene Ontology Consortium : Gene Ontology: tool for the unification of biology. In *Nature Genet*, 2000, 25, pp. 25-29
19. N.J. Mulder and R. Apweiler and T.K. Attwood and A. Bairoch and A. Bateman and D. Binns and P. Bradley and P. Bork and P. Bucher and L. Cerutti and R. Copley and E. Courcelle and U. Das and R. Durbin and W. Fleischmann and J. Gough and D. Haft and N. Harte and N. Hulo and D. Kahn and A. Kanapin and M. Krestyaninova and D. Lonsdale and R. Lopez and I. Letunic and M. Madera and J. Maslen and J. McDowall and A. Mitchell and A.N. Nikolskaya and S. Orchard and M. Pagni and C.P. Ponting and E. Quevillon and J. Selengut and C.J. Sigrist and V. Silventoinen and D.J. Studholme and R. Vaughan and C.H. Wu : InterPro, progress and status in 2005. In *Nucleic Acids Res*, 2005, 33, pp. D201-205
20. D. Gilbert : Biomolecular interaction network database. In *Brief Bioinform*, 2005, 6, pp. 194-198
21. Reiter, R. : Towards a Logical Reconstruction of Relational Database Theory. In Brodie, M., Mylopoulos, J., Schmidt, J. (eds.), *On Conceptual Modelling*, Springer-Verlag, 1984, pp. 191-233.
22. Meersman, R. : The Use of Lexicons and Other Computer-Linguistic Tools in Semantics, Design and Cooperation of Database Systems. In Zhang, Y., Rusinkiewicz, M., Kambayashi, Y. (eds.), *Proceedings of the Conference on Cooperative Database Systems (CODAS 99)*, Springer-Verlag, 1999, pp. 1-14.
23. Meersman, R. : Ontologies and Databases: More than a Fleeting Resemblance. In d'Atri, A., Missikoff, M. (eds.), *OES/SEO 2001 Rome Workshop*, Luiss Publications.
24. Spyns, P., Meersman, R. and Jarrar, M. : Data Modelling versus Ontology Engineering. *SIGMOD Record: Special Issue on Semantic Web and Data Management*, 2002, 31(4), pp. 12-17.
25. Guarino, N., and Giaretta, P. : Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Mars, N. (ed.) *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, IOS Press, Amsterdam, pp. 25-32.
26. A. Kasprzyk and D. Keefe and D. Smedley and D. London and W. Spooner and C. Melsopp and M. Hammond and P. Rocca-Serra and T. Cox and E. Birney : EnsMart: a generic system for fast and flexible access to biological data. In *Genome Res*, 2004, 14, pp. 160-169
27. D. Maglott and J. Ostell and K.D. Pruitt and T. Tatusova : Entrez Gene: gene-centered information at NCBI. In *Nucleic Acids Res*, 2005, 33, pp. 54-58
28. Meersman, R. : Semantic Web and Ontologies: Playtime or Business at the Last Frontier in Computing? In *NSF-EU Workshop on Database and Information Systems Research for Semantic Web and Enterprises*, 2002, pp. 61-67.
29. P. De Leenheer and R. Meersman : Towards a formal foundation of DOGMA ontology: part I. Technical Report STAR-2005-06, VUB STARLab, 2005.
30. De Leenheer, P. and de Moor, A. and Meersman, R. : Context Dependency Management in Ontology Engineering. Technical Report STARLab, Brussel, 2006.

31. Deray T. and Verheyden P. : Towards a semantic integration of medical relational databases by using ontologies: a case study. In, R. Meersman, Z. Tari et al.,(eds.), *On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops*, LNCS 2889, pp. 137 - 150, 2003. Springer Verlag.
32. Verheyden P. Deray T. and Meersman R., *Semantic Mapping of Large and Complex Databases to Ontologies: Methods and Tools*. Technical Report 25, STAR Lab, Brussel, 2004.
33. Verheyden P., De Bo J. and Meersman R. : Semantically unlocking database content through ontology-based mediation . In, Bussler C., Tannen V. and Fundulaki I.,(eds.), *Proceedings of the 2nd Workshop on Semantic Web and Databases (SWDB 2004)*, LNCS 3372, pp. 109 - 126, 2005. Springer Verlag.
34. Trog D. and Vereecken J. : *Context-driven Visualization for Ontology Engineering*. Master thesis, Vrije Universiteit Brussel, 2006
35. Halpin T. : *Information Modeling and Relational Databases*. Morgan Kaufmann Publishers Inc. , 2001
36. Barriot, R and Poix, J and Groppi, A and Barré, A and Goffard, N and Sherman, D and Dutour, I and de Daruvar, A : New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. In *Nucleic Acids Res*, 2004, 32(12), pp. 3581-3589

Bringing Together Structured and Unstructured Sources: The OUMSUIS Approach

Gayo Diallo, Michel Simonet, and Ana Simonet

Laboratoire TIMC-IMAG, UJF Grenoble,
Fac De Médecine 38700 La Tronche, France
`firstname.lastname@imag.fr`

Abstract. Integration of heterogeneous sources is a means to offer the user an access to multiple information sources in a unified way through queries submitted to a global schema. We propose a semantic web-based mediator model, to provide unified access to various sources which may be both structured (database systems) and unstructured (textual medical reports, scientific publications, etc.). The mediator level is composed of the global ontology and a set of ontologies which make it possible to characterize sources (one or several ontologies by source). Unstructured sources are seen through their Semantic Document Representation obtained by a semantic characterization process. A reverse engineering process is then applied on each Semantic Document Representation schema and each structured source schema in order to provide semi-automatically a set of local ontologies. These local ontologies are articulated around the global schema following the global centric approach. A service called Terminology Server (ServO) is used to perform queries and manage the ontologies. The ontology-based query model combines databases and information retrieval techniques. We illustrate this approach with a case study in the brain disease field but it is sufficiently generic to be used in other domains.

1 Motivation

The increase of the number of information sources requires efficient and flexible frameworks for their integration. In a recent past organizations and companies often manipulated information provided by (relational) database systems, but nowadays most available information resources are in textual form – including technical documents, scientific literature, on-line web pages, etc. Database systems (structured sources) are usually well structured and thus make available to the users efficient and accurate access to a large collection of information through data retrieval techniques (exact search). Textual data sources (qualified as informal or unstructured sources) are accessed through Information Retrieval (IR) techniques (usually qualified as approximate search) [16]. Accessing multiple sources in order to satisfy a user need is particularly useful in medicine.

Indeed, it may be necessary to query medical report repositories, on-line medical literature (e.g., Pubmed¹), patient database systems, etc., all this through considerable manual effort.

Generally speaking, the integration of sources is an area of computer science related to information exchange and information gathering from disparate and possibly heterogeneous information sources. It frees the users from tasks such as finding the relevant data sources, interacting with each source in isolation, and selecting, cleaning, and combining data from multiples sources [18]. The multiplication of sources may bring a certain heterogeneity either structural (schema heterogeneity) or semantic (data heterogeneity) [17]. Structural heterogeneity means that different information systems store their data in different structures. Semantic heterogeneity considers the contents of information items and their intended meaning.

The work presented in this paper describes the OUMSUIS (Ontology-based Unified Management of Structured and Unstructured Information Sources) approach. Ontologies, as an explicit specification of a conceptualization [13], can be used to solve the heterogeneity problems since they represent explicitly and semantically the content of sources. In medicine, the targeted application domain of OUMSUIS, substantial work has been done to develop standards, medical terminologies and coding systems (SNOMED², MeSH³ and UMLS, which integrates more than 100 of the most relevant vocabulary sources in medicine [2]). Those standards may be exploited to build or to enrich ontologies which help characterizing sources for integration purpose.

In the OUMSUIS system we consider that integration is performed according to a particular need. Thus from the same set of sources, several integrated views may be provided. Our contribution in this paper mainly consists of a general framework for unifying the management of both structured and unstructured information sources through ontologies expressed in OWL (Ontology Web Language)⁴. OWL is a recommendation of the W3C to represent ontologies. In this paper we focus on the general foundations of the approach, which will be presented as follows. Section 2 describes the case study in the brain disease area which we use throughout the paper. Section 3 describes the overview of the OUMSUIS approach. We describe the mapping expressions and query service for OUMSUIS in Section 4. Section 5 describes related work in the information integration area in general and particularly in medicine. We conclude and present our forthcoming work in Section 6.

2 Case Study in the Brain Disease Field

We first introduce a case study in the brain disease field which we will use throughout the paper to illustrate our approach. The neuropsychology center

¹ www.pubmed.gov

² <http://www.snomed.org/>

³ <http://www.nlm.nih.gov/mesh/>

⁴ <http://www.w3.org/TR/owl-guide/>

of the Pitié-Salpêtrière hospital in Paris is specialized in the assessment of the neuropsychological after-effects of cerebral lesions. In the center they collect and process neuroanatomical, neuropsychological and behavioral data of cerebral-injured patients. They establish neuropsychological reports based on observations and objective tests, i.e., neuropsychological tests chosen by the doctor according to the symptoms of the patient. A neuropsychological report describes the cognitive functions evaluated (e.g., memory, language) through a list of cognitive tests.

A neuropsychological report is not intended only to quantify and qualify general cognitive functions. It helps understanding their different components. For example the *Memory* function is evaluated through its subdivisions such as semantic memory, procedural memory, etc. In the same manner *language* may be evaluated through its verbal and written modalities. The center manages several kinds of data: neuropsychological reports, multimedia documents representing MRI scan of the patients brain used to localize lesions, administrative information about the patients and staff, etc. In order to get more information about a given brain disease and to be informed about the state of the art in the domain, they accesses some medical literature repositories such as Pubmed. The information system of our target application domain must answer queries like:

- Documents dealing with given concepts of an ontology
- Doctor who has examined a given patient
- Test value obtained by a patient after a neuropsychological examination
- MRI scan showing a set of affected anatomical zones, etc.

Table 1. Patient Relational Database. Information concerning the doctors (Table *Medecin*), the various neuroanatomical examinations of the patients (Tables *ExamPatient* and *LesionPatient*) is managed by this database. The *IntituleZoneAnat* attribute from the table *LesionPatient* denotes a brain anatomical part affected by a lesion.

ExamPatient (IDExamPat, IDCAC, ExamDate, ExamCode, IDMed, noteMede)
Examen (codeExam, LibExam, typeExam)
LesionPatient (IDCAC, IntituleZoneAnat)
Medecin (IDmedecin, NomMed, PnomMed, ADRMed, VilleMed, TelMed, SpecialitMed)
Patient (IDCAC, NumDossier, PnomPatient, Sexe, AnneeNais, lateralite, VillePatient)

3 Overview of the OUMSUIS Framework

OUMSUIS follows a virtual centric view (also called Global As View), providing a unified view on the different data sources manipulated by an organization through a global schema. The information sources differ by their structure, format and contents. OUMSUIS deals explicitly with the integration of separate texts collection with relational database systems. The integration process that we propose is based on four principles.

1. Each informal source to be integrated is characterized by its own ontologies (semantic indexing), which can be shared by other sources. The ontologies themselves are not stored in the source but are referenced through their URI (Uniform Resource Identifier).
2. Each source (with its schema) is represented by a source ontology which has to be articulated with the global called ontology for integration perspective (*OIP*). The *OIP* is obtained from the domain ontology (which already exists or is built according to the need of the organization) and all the local ontologies.
3. The global level includes the *OIP* and the set of ontologies used to characterize the sources. We distinguish ontologies helping to characterize local sources from the *OIP*.
4. Each source may be accessed autonomously if needed and an administrator is in charge of the system management.

The OUMSUIS integration system, following the model presented in [5], is a quadruple $\{OIP, S_{local}, \mathcal{O}, \mathcal{M}\}$ where

- *OIP* is the global ontology for integration perspective (over an alphabet A_G),
- S_{local} (over an alphabet A_L) is set of local ontologies built on top of local schemas ($S_{local} = S_{structured} \cup S_{unstructured}$),
- $\mathcal{O} = \mathcal{O}_i^p_{i=1}$, is the set of ontologies characterizing local sources.
- \mathcal{M} is the mapping between *OIP* and S_{local} .

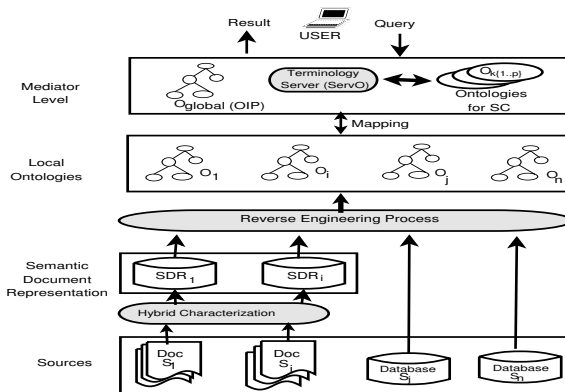


Fig. 1. OUMSUIS Layers Representation

Figure 1 presents the different layers of OUMSUIS. The bottom layer represents the sources to be integrated. An intermediate layer represents the *SDR* of unstructured sources (currently a PostgreSQL database). The next layer

represents local ontologies obtained through a reverse engineering process. We have then the global level (mediator) which has both ontologies for local sources characterization (SC) and the *OIP* denoted as O_{global} . OUMSUIS takes into account that several ontologies can provide different perspectives on the same source. A possibility is given at the mediator level to change the user perception by using two primitives: *AssignOnto* and *RemoveOnto*, respectively to add and to remove an ontology (see Section 4).

3.1 Ontologies Involved in the Integration Process

Several kind of ontologies are involved in the OUMSUIS integration process.

Ontology for Integration Purpose. The ontology for integration purpose (*OIP*) acts as the global schema and is used to query the system. Currently it is built manually from the domain ontology of the organization and all the local ontologies obtained during the reverse engineering process. An *OIP* is built for each integration need.

Ontologies for Semantic Sources Characterization. For each unstructured source the automatic semantic characterization process uses a set of ontologies and optionally a specific ontology for named entities recognition (NER). The former contains concepts and relations between concepts describing important notions in a given domain (e.g., ontology of brain cognitive function, ontology of brain anatomy, etc.). This kind of ontology may be a terminological ontology (TO) according to Sowa who defines a terminological ontology as an ontology whose categories need not be fully specified by axioms and definitions⁵. The latter is used to classify named entities identified in textual documents. In our case classical named entities such as *Person*, *Location* and *Organization*⁶ are complemented by number expressions representing cognitive test values. Indeed in the brain disease field, it is important to identify result values obtained by a patient after a set of neuropsychological cognitive tests.

We have chosen the OWL language to represent the ontologies. Its sub-language OWL Lite supports those users primarily needing a classification hierarchy and simple constraints, and it provides a quick migration path for thesauri and other taxonomies [1]. For the brain disease case study, an anatomical ontology has been built semi-automatically from the *neuronames* brain hierarchy developed by Bowden and Martin [3] and UMLS using one of the OUMSUIS module. We have also developed manually a brain function ontology edited under the Protégé editor⁷. Below is an extract (DL notation) of the description of the concept *SemanticMemory* (ontology of the brain functions) denoted by the terms *Mémoire Sémantique* in French and *Semantic Memory* in English. Its alternative terms are not included.

⁵ <http://www.jfsowa.com/ontology/gloss.htm>

⁶ http://www.itl.nist.gov/iaui/894.02/related_projects/muc/

⁷ <http://protege.stanford.edu/>

3.2 Semantic Document Representation

The first level of homogeneity is obtained by a semantic characterization process which provides a repository named *SDR*, based on a relational database implemented currently with the DBMS PostgreSQL. The *SDR* provides a hybrid representation of an unstructured source: term/concept indexing, cataloguing information, classified named entities. From it, several outputs may be provided according to the translation modules which have been developed. Thus it is possible to have an RDF representation of the content of the *SDR*, an XML schema-based representation and OWL ontology-based representation. Three modules are used to complete the semantic characterization process.

- The HybridIndex module, which automatically builds a hybrid index using the ontologies declared for the source [4]. It provides weighted vectors of terms and concepts for each document of the source.
- The Metadata module which loads the document Dublin Core cataloguing⁸ information provided by the Annotation Tool [15] and corrects the errors which may occur during the automatic characterization process.
- Finally the module for named entity recognition is used to recognize named entities (cognitive test values) in the source using the ontology for NER; user of this module is optional depending on the source to be treated. It is less pertinent to identify cognitive test values in medical literature than in neuropsychological reports.

4 Mapping Expression and OUMSUIS Querying

The global-local ontology mapping \mathcal{M} is an important part of the infrastructure since it specifies how the concepts and relations in the global ontology and the local ontologies are articulated.

4.1 Global and Local Ontologies Articulations

The global schema is represented by the *OIP* and each local source is viewed through a local ontology obtained semi-automatically. Articulations between the global and local ontologies are expressed by axioms which represent how to obtain global concepts and/or relations. The articulations follow the global centric approach where the concepts/relations of the global ontology are expressed as queries over the local ontologies.

4.2 Query Service for OUMSUIS

AssignOnto and RemoveOnto We may consider during the querying phase that for a particular source the ontologies for semantic characterization may change over time. A source in our approach does not need to store the ontologies helping to characterize it, since different sources may be characterized

⁸ <http://dublincore.org/>

by the same ontology. Moreover, storing the different ontologies in the sources themselves is space-consuming and ontologies might be duplicated in several sources. We have defined two primitives:

- AssignOnto(S , $Onto$): which informs the mediator that the informal source S is characterized by the ontology $Onto$. During the querying process the system examines ontologies currently being allowed for each source.
- RemoveOnto(S , $Onto$): this operation performs the inverse of AssignOnto. It removes a given ontology for a source S . For example, the administrator of the system may decide to remove the possibility for an informal source S to be viewed by a given ontology during its maintenance process.

Terminology Server: The ServO Module. In section 3.1 we show that several characterization ontologies are simultaneously manipulated within the OUMSUIS system. When a term expressed in natural language is provided by the user for querying purposes, it has to be associated with the concept it represents. We have developed a module called ServO which provides an index of ontologies. It is based on the Apache Lucene API⁹ which offers full-text search capabilities. The ServO module automatically looks for links between a given term and the concepts of the ontologies. Moreover it is possible to ask for concepts having certain declared properties. The result is a ranked list (for each ontology) of concepts related to the term. The retrieved concepts may be used later to query the OUMSUIS system. The ontologies exploited by the ServO module may vary over time according to AssignOnto and RemoveOnto primitives.

System Querying. Queries to the OUMSUIS mediator are expressed by a query language over the alphabet A_{global} and are expressed in the OWL-QL query language¹⁰. A query is processed by means of unfolding, i.e., by expanding each atom according to its associated definition in the mapping \mathcal{M} , so as to come up with sources ontology atoms. For example if we assume that $OIP:HasConcept$ denotes the *HasConcept* property of the *OIP*, the simple query "Retrieve items related to *Parietal Lobe*" expressed as $(?x \text{ OIP:HasConcept ParietalLobe must-bind } ?x)$ will return both patients affected by a lesion in the parietal lobe and documents indexed by the same concept or any of its sub concepts.

5 Related Work

Two major approaches for information integration are commonly used: (1) the materialized approach (also called data warehouse) and (2) the virtual approach (mediator-based). In the materialized approach actual data are duplicated in a central repository while in the latter approach, the actual data resides in the sources, and queries against the integrated view are decomposed into subqueries posed to the sources. Ontologies have been widely used by systems developed

⁹ <http://lucene.apache.org/>

¹⁰ ksl.stanford.edu/projects/owl-ql/

for information integration purpose as reported in the literature [6][7]. A recent survey of ontology-based approach may be found in [10][11].

From the point of view of mapping expression (Global As View) OUMSUIS is close to TSIMMIS [12] and MOMIS [19] systems while from the point of view of multiple ontologies in the global level it may be compared to OBSERVER [7]. We deal with sources containing sometimes different (but related through ontologies) information while the problems commonly tackled by the systems above (schema matching, conflicts resolving, mismatches, etc.) are heterogeneities between different representations of essentially the same – or very similar – real world entities [21].

In the life sciences domain work on ontology-based integrating heterogeneous data sources includes TAMBIS [8] and ONTOFUSION [9]. TAMBIS is an ontology-centred system querying multiple heterogeneous bioinformatics sources. Only one general ontology is allowed at any given time. ONTOFUSION is another ontology-based system designed for biomedical database integration. ONTOFUSION takes into account only structured sources (UMLS, GO, etc.); the virtual schemas are manually built and only direct mappings between relational tables and concepts of an ontology are provided. Services offered by those systems may be compared to those offered by the ServO module described above. We can also mention the Sequence Retrieval System (SRS) [20] which is closer to a keyword-based retrieval system than an integration system. OUMSUIS do not allows only accessing multiple biomedical vocabularies or ontologies; it allows accessing (multilingual) resources semantically characterized by those vocabularies or ontologies. Moreover it distinguishes the ontologies used for semantic characterization purpose from the ontologies representing both the sources and the global level. This distinction which is not commonly made minimizes the impact of the evolution of ontologies on the system.

6 Conclusion

In this paper we have presented the OUMSUIS approach for structured and unstructured information integration and a case study in the brain disease field. Our approach uses the global centric approach to express mappings between global and local schemas through a semantic web language. OUMSUIS distinguishes itself by its separation of ontologies helping to characterize a given source and the domain ontology. As an ontology may characterize several sources, managing its persistence in the source itself is not a suitable way; so we propose to keep it in one location identified by an URI. At the mediator level the user's views of sources may change depending on the current ontologies. In order to interact efficiently with the different ontologies, OUMSUIS has an ontologies server module to help finding concepts from user natural language terms. The development of OUMSUIS is an ongoing work. Further considerations including queries rewriting, sources evolution, overlapping of concepts from different characterization ontologies had to be considered.

References

1. Antoniou G., van Harmelen, F.: Web Ontology Language. In Handbook on Ontologies, S. Staab, R. Studer (Eds.) Springer-Verlag, Berlin Heidelberg New York (2004)
2. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, vol 32 (Database issue), (2004) 267–270
3. Bowden D.M., Martin R.F.: Neuronames brain hierarchy. *Neuroimage* **2** (1995) 63–83
4. Diallo G., Simonet M., Simonet A.: An Approach of Automatic Semantic Annotation of Biomedical Texts. In Proceedings of IEA/AIE'06, LNAI 4031, Springer-Verlag (2006) 1024–1033
5. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., Rosati, R.: Knowledge representation approach to information integration. In Proc. Of AAAI Workshop on AI and Information Integration, pages . AAAI Press/The MIT Press, (1998) 58–65
6. Köhler, J., Philippi, S., Lange, M.: SEMEDA: ontology based semantic integration of biological databases, *Bioinformatics* 19 (18) (2003) 2420–2427.
7. Mena, E., Illarramendi, A., Kashyap, V., Sheth, A.: OBSERVER: An approach for query processing. In: global information systems based on interoperation across pre-existing ontologies. In: International journal on Distributed And Parallel Databases (DAPD), ISSN 0926-8782, Vol. 8, No.2, April (2000) 223–271
8. Paton, N. W., Stevens, R., Baker, P., Goble, C., Bechhofer, S., Brass, A.: Query Processing in the TAMBIS Bioinformatics Source Integration System. *SSDBM* (1999):138-147
9. Pérez-Rey, D., Maojo, V., Garcia-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martin-Sanchez, F., Sousa, A.: ONTOFUSION: Ontology-based integration of genomic and clinical databases, In *Computers in Biology and Medicine*, vol ? (2005)
10. Wache, H., Vögel, T., Visser U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hbner, S.: Ontology-based Integration of Information- A survey of existing approaches. In Proceedings of IJCAI01 Workshop on Ontologies and Information Sharing, Seattle, Washington (2001)
11. Doan A., Halevy A. Y.: Semantic Integration Research in the Database Community: A Brief Survey. *AI Magazine*, Special Issue on Semantic Integration, Spring (2005)
12. Garcia-Molina, H., Hammer J., Ireland, K., Papakonstantinou, Y., Ullman, J., Windom, J.: Integrating and Accessing Heterogenous Information Sources in TSIMMIS. In Proceedings of the AAAI Symposium on Information Gathering, Stanford, California, March (1995) 61–64
13. Gruber, T. R. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, (1993)
14. Simonet, A., Simonet M.: The ISIS Methodology for Object Database Conceptual Modelling, Poster E/R 99, 18th Conference on Conceptual Modeling, Paris, 15-18 Nov. (1999)
15. Patriarche R., Gedzelman S., Diallo G., Bernhard D., Bassolet C.G., Ferriol S., Girard A., Mouries M., Palmer P., Simonet M.: A Tool for Textual and Conceptual Annotation of Documents. In Proceedings of E-Challenge2005, Ljubljana, Slovenia, (2005)
16. van Rijsbergen, C., J.: Information retrieval. London, Butterworths, (1979)

17. Kim, W., Seo, J.: Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 24(12), (1991) 12–18
18. Borgida, A., Lenzerini, M., Tosati, R.: *Description Logics for Databases*. In *The Description Logics Handbook, Theory, Implementation and Applications*, Cambridge University Press (2002)
19. Beneventano, D., Bergamaschi S.: The Momis methodology for integrating heterogeneous data sources. *IFIP2004 Congress Topical Sessions*, Toulouse, France (2004) 19–24
20. Zdobnov, E. M., Lopez, R., Apweiler, R., Eitzold, T.: The EBI SRS server - recent developments. *Bioinformatics*, Vol. 18, Num. 2, pp. 368-373, 2002.
21. Rahm, E., Bernstein, P., A.: A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, vol 10, No.4 pp 334-350, (2001)

Reactome – A Knowledgebase of Biological Pathways

Esther Schmidt, Ewan Birney, David Croft, Bernard de Bono,
Peter D'Eustachio, Marc Gillespie, Gopal Gopinath, Bijay Jassal,
Suzanna Lewis, Lisa Matthews, Lincoln Stein,
Imre Vastrik, and Guanming Wu

European Bioinformatics Institute
(EMBL-EBI), Wellcome Trust Genome
Campus, Hinxton, Cambridgeshire,
CB10 1SD, UK
{eschmidt, birney, croft, bdb, bj1,
vastrik}@ebi.ac.uk
Cold Spring Harbor Laboratory,
Cold Spring Harbor, New York 11724, USA
{eustachi, gillespm, gopinath,
lisa.matthews, lstein, wugm}@cshl.edu
Department of Molecular and Cell Biology,
University of California
Berkeley, Berkeley, California, USA
suzi@fruitfly.org

Abstract. Reactome (www.reactome.org) is a curated database describing very diverse biological processes in a computationally accessible format. The data is provided by experts in the field and subject to a peer review process. The core unit of the Reactome data model is the reaction. The entities participating in reactions form a network of biological interactions. Reactions are grouped into pathways. Reactome data are cross-referenced to a wide selection of publically available databases (such as UniProt, Ensembl, GO, PubMed), facilitating overall integration of biological data. In addition to the manually curated, mainly human reactions, electronically inferred reactions to a wide range of other species, are presented on the website. All Reactome reactions are displayed as arrows on a *Reactionmap*. The *Skypainter* tool allows visualisation of user-supplied data by colouring the *Reactionmap*. Reactome data are freely available and can be downloaded in a number of formats.

Keywords: knowledgebase, pathways, biological processes.

1 Introduction

The Human Genome Project has provided us with huge amounts of data, including a good approximation of the encoded components that make up a living cell [1]. This was an important step, but now another challenge is the question how all these components actually interact in a cell, and how they bring about the many processes essential for life. A lot of these individual processes have already been studied extensively, but the

resulting information is spread throughout the scientific literature. There is an urgent need to bring this information together, allowing the scientist to see connections and understand dependencies in this complex network of interacting entities.

The Reactome database has been developed to provide such a platform, not only to collect information in one place, but also to present it in a systematic, computationally accessible manner. An important aspect of this project is the free availability of all data, and the integration with other publically available databases in order to enhance the potential for the user to extract as much relevant information as possible about a process of interest.

2 Contents and Curation Process

Reactome is aiming at comprehensive coverage of human cellular processes, be it metabolic reactions, catalyzed by an enzyme, or be it complex formation, transport across a membrane, DNA repair or signal transduction. At present, topics included are apoptosis, cell cycle, transcription, mRNA processing, translation, post-translational modification, signalling pathways (insulin, notch), hemostasis, energy metabolism (TCA cycle, glycolysis), amino acid, lipid and nucleotide metabolism as well as xenobiotic metabolism (see Fig.1).

Reactome is a manually curated database. Data are obtained directly from experts. Suitable topics for inclusion in Reactome are identified and independent researchers who are recognized in the field are approached. A Reactome curator and the expert then work together to agree on an outline and structure the data to conform to the Reactome data model. A major emphasis is put on confirming the exact identity of the entities involved by assigning the appropriate identifiers from UniProt [2] or ChEBI (www.ebi.ac.uk/chebi/), respectively. All reactions need to be backed up by a literature reference, stating the Pubmed identifier whenever available. The curator also makes sure that appropriate Gene Ontology (GO) [3] terms are cross-referenced for catalytic activities, cellular locations and biological processes. This curation process is followed by an internal review to ensure consistency, and peer review by another expert familiar with the topic.

3 Website

The front page of the Reactome website (Fig.1) displays a *Reactionmap*, followed by a section listing the high-level topics represented in Reactome, and a section giving some general information on the project including latest news. In the *Reactionmap* each reaction is represented as an arrow. The topics are arranged as distinct patterns in the map so that the user can easily identify his area of interest. A separate *Reactionmap* can be displayed for each species, providing a quick overview as to which pathways are present or absent in a given species. Mousing over the *Reactionmap* or the topic section highlights the corresponding area in the other section, and both can serve as entry points into the detailed content pages of Reactome.

The detailed content pages (Fig.2) present the event hierarchy, display diagrams indicating the entities involved in the event as well as preceding and following events, and are often accompanied by an author-supplied illustration. A textual description of the event, literature references as well as species and compartment details are given. Orthologous events in other species and relevant GO biological process terms are found here as well. Physical entities are given along with relevant links to publically available databases such as UniProt, Ensembl [4], KEGG [5], ChEBI, etc. Internal links take the user to more detailed pages on the entity in question, for example giving information on the composition of a complex.

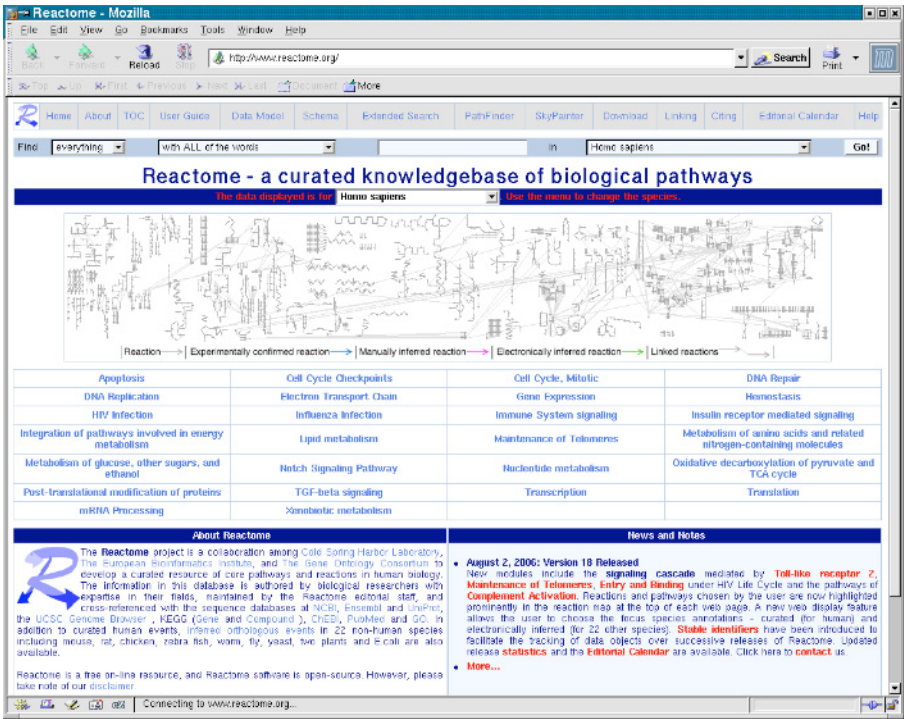


Fig. 1. The Reactome front page. The list of topics in Reactome and the Reactionmap are shown for human by default. Electronically inferred events for 22 other species can be displayed by choosing the species from the drop down menu above the *Reactionmap*.

Other features of the website include a *User Guide*, information on the data model as well as citing and linking to Reactome, and an editorial calendar, indicating topics planned for inclusion in the future. Simple searches can be performed on every content page, and there is an extended search option for users familiar with the data model.

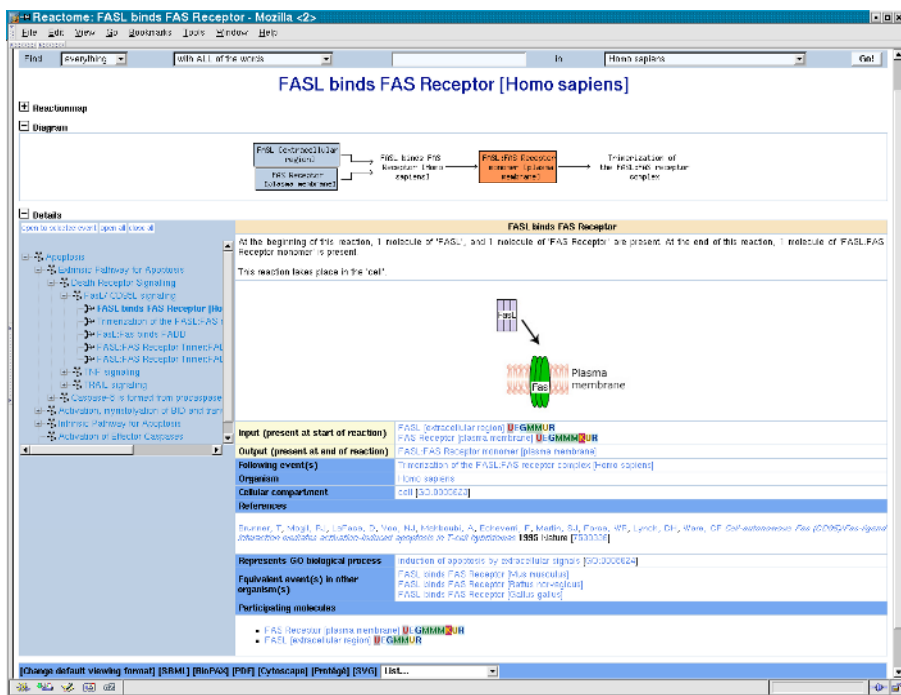


Fig. 2. A detailed content page



Fig. 3. A Skypainter result page for time series data

The *Skypainter* tool (Fig.3) gives the user the option of submitting a list of identifiers, for example from a microarray experiment, in order to visualise reactions that match these identifiers on the *Reactionmap*. A statistical analysis based on the hypergeometric test (http://en.wikipedia.org/wiki/Hypergeometric_distribution), is performed to colourise pathways according to the statistical likelihood that they would contain the listed genes by chance. This highlights those pathways in which the uploaded genes are overrepresented. A large number of gene identifiers, including EntrezGene names, accession numbers and Affymetrix probe sets can be recognised by the *Skypainter*. It also accepts numeric values, such as expression levels from a microarray experiment. For example, a researcher who is using a microarray to compare a cancerous tissue to a normal control can upload the intensity values from the two experiments to the *Skypainter*, and it will colourise the *Reactionmap* with red and green to indicate which reactions have genes that are increased or decreased in the malignant cells relative to the normal controls. When submitting time series data, the changes across the *Reactionmap* can be displayed in an animation, showing the coloured *Reactionmaps* in succession. Reactions hit by an identifier are listed below the reaction map, providing links to the corresponding detailed content pages.

4 Data Model

The Reactome data model has been developed with the view to allow representation of a variety of different cellular processes: In metabolic reactions, a chemical entity is transformed into a different chemical entity by an enzyme acting as catalyst; molecules are transported from one cellular compartment into another across membranes; proteins undergo posttranslational modifications, or form complexes; a protein, modified in one reaction, may act as a catalyst in a following reaction; a chemical entity, produced by a metabolic reaction, may be required as an activator of a reaction in a signaling cascade. Thus, the data model needs to be able not only to describe individual reactions; it needs to deal with different kinds of reactions in a consistent manner so that the relationships and interactions of the entities in this network can be expressed.

Reactome uses a frame-based knowledge representation. Concepts (such as reaction, complex, regulation) are expressed in classes, and the knowledge is represented as instances of these classes. Classes have attributes attached that hold properties of the instances, e.g. the identity of entities participating in a reaction.

One of the main classes of the Reactome data model is the *Reaction*. Its main attributes are *input*, *output* (both accepting instances of the *PhysicalEntity* class) and *catalystActivity* (accepting an instance of the *CatalystActivity* class, which in turn holds a Gene Ontology molecular function term under *activity* and an instance of the *PhysicalEntity* class under the *physicalEntity* attribute). Other attributes include *compartment*, which holds a Gene Ontology cellular component term to indicate the cellular location of the reaction, *literatureReference*, *summation*, *figure*, *species*, *goBiologicalProcess* and *precedingEvent*. Related reactions in other species are attached via the *orthologousEvent* and *inferredFrom* attributes. Reactions are grouped into *Pathways*, which can again be components of higher-level pathways.

Another important class is the *PhysicalEntity*. This class is subdivided into *GenomeEncodedEntity* to hold species-specific molecules, *SimpleEntity* for other chemical entities such as ATP, *Complex* and *Polymer* for entities with more than one component, and *EntitySet* for groups of entities that can function interchangeably in a given context. Post-translational modifications are expressed through the *hasModifiedResidue* attribute, which holds an instance of the class *ModifiedResidue* that in turn contains attributes describing the nature of the modification. Modified proteins are treated as distinct entities from unmodified proteins, and molecules in one cellular compartment are distinct entities from molecules in another compartment.

Such *PhysicalEntity* instances that represent the same chemical entity in different compartments, or different modified forms of the same protein, share numerous invariant features such as names, molecular structure and links to external databases like UniProt or ChEBI. To enable storage of this shared information in a single place, and to create an explicit link among all the variant forms of what can also be seen as a single chemical entity, Reactome creates instances of the separate *ReferenceEntity* class. A *ReferenceEntity* instance captures the invariant features of a molecule. A *PhysicalEntity* instance is then the combination of a *ReferenceEntity* attribute (e.g., Glycogen phosphorylase UniProt:P06737) and attributes giving specific conditional information (e.g., localization to the cytosol and phosphorylation on serine residue 14).

5 Quality Assurance

For any database, data consistency is of utmost importance. In order to provide reliable data, a series of quality assurance tests is performed before data is publicly released in Reactome. The main features of this procedure are a check for the presence of essential attributes that are mandatory for a given class, as well as a check for imbalances between input and output protein entities. The latter is based on the principle that proteins that act as input for a reaction need to be present in some (possibly modified) form in the output of the reaction as well, except in synthesis or degradation reactions. Other checks ensure consistency in terms of cellular compartments or species origin for the entities involved. These automated checks are performed in addition to the external and internal review processes mentioned above.

6 Orthology Inference

The main focus of Reactome is manual curation of human biological processes. In some cases, the actual experimental evidence for a biological reaction has been demonstrated in another species and the occurrence of this reaction in human can only be inferred by the experts. Reactome deals with this scenario by describing the reaction in the other species, backed up by a literature reference. A human reaction is then described as well, pointing to the reaction in the other species via the *inferredFrom* relationship. Thus the evidence can always be tracked back to the original experiment.

In addition to these manually curated events in other species, Reactome also provides electronically inferred reactions. All human reactions that involve at least one protein with known sequence and are not themselves inferred from the other species under consideration, are eligible for orthology inference. Eligible reactions are submitted to an automated protocol, attempting inference to 22 other species. The rationale for orthology inference is that if all proteins involved in a human reaction have an orthologous protein in the other species, the reaction is likely to occur in the other species as well. A Reactome reaction is then created for the other species, with all species-unspecific entities copied over, and species-specific entities replaced by the respective orthologous entities. Such reactions are marked as electronically inferred reactions and point to the manually curated human reaction via the *inferredFrom* relationship.

Orthology relationships between proteins are obtained from the OrthoMCL database, which provides orthologue groups based on sequence similarity clustering [6] [7].

When applying the strict inference criteria where each protein needs to have an orthologue, reactions involving large complexes often get excluded from inference. To allow for inter-species differences in complex composition, the criteria are relaxed for complexes such that reactions are inferred to the other species when at least 75% of the protein components in a complex have orthologues.

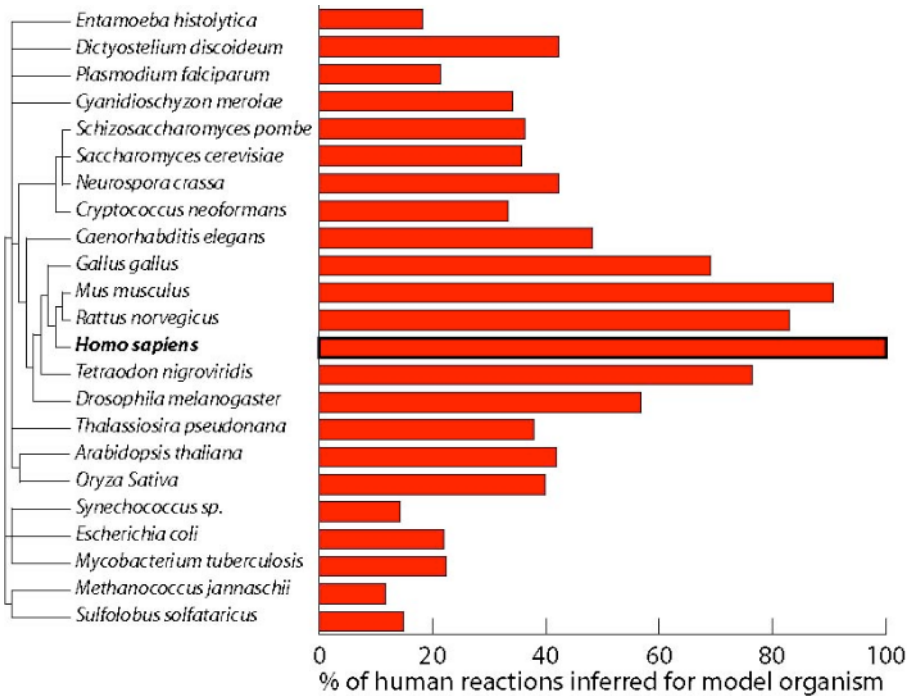


Fig. 4. Result of orthology inference. The percentage of eligible human reactions inferred to each species is given. The total number of eligible human reactions in release 18 was 1590.

The species included in the Reactome inference procedure cover a wide range of phylogenetic groups, and include model organisms such as mouse, drosophila and C.elegans. For mouse, 1442 out of 1590 (91%) of eligible human reactions were inferred for Reactome release 18, while 904 out of 1590 (57%) of eligible reactions were inferred to drosophila and 348 out of 1590 (22%) to E.coli. Fig. 4 shows a graph with inference figures for all species.

Obviously, such electronic predictions need to be considered with caution as sequence similarities don't necessarily imply functional equivalence. However, they can serve both as a starting point for manual curation in these species and as entry points into the database via protein identifiers from non-human species.

7 Downloads

All Reactome data and software are available free of charge to all users. Various download formats are available via the website. SBML [8] and BioPAX (<http://www.biopax.org/index.html>) are exchange formats for systems biology and

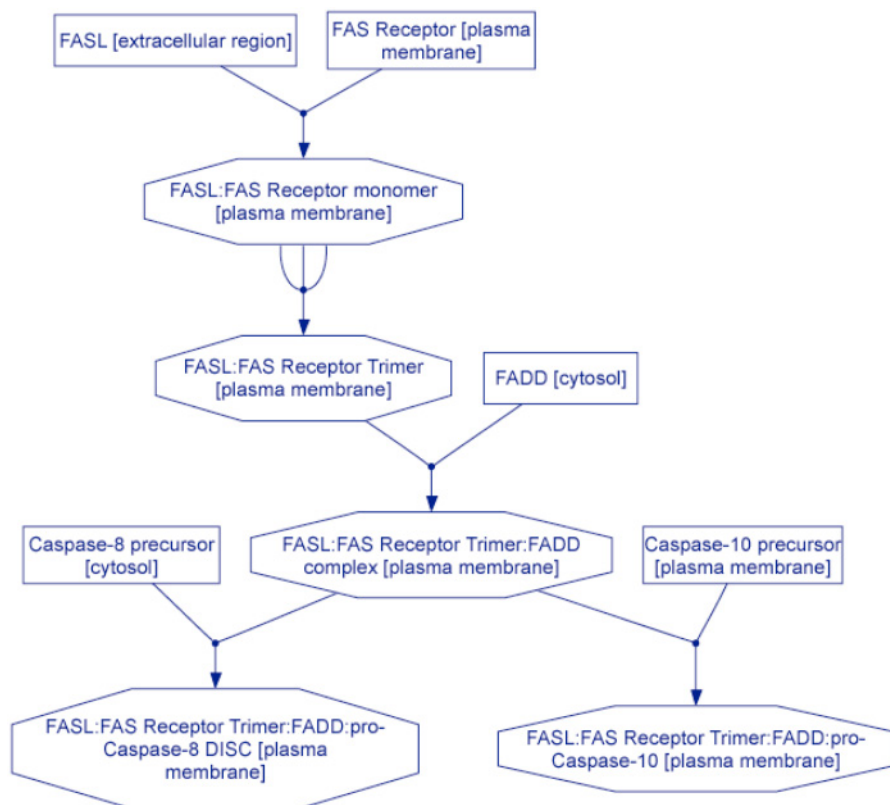


Fig. 5. Pathway diagram in svg format, generated for FasL/CD95L signaling

pathway databases, respectively. Textual description of pathways together with illustrations can also be downloaded in pdf format. Programmatically generated pathway diagrams can be saved in scalable vector graphics (SVG) format (Fig.5). Various lists can be obtained, e.g. a list of all protein identifiers involved in a pathway or reaction. The entire database is available as a mysql dump, and the website can be installed locally. Downloads for the data entry tools used by authors and curators are also available.

8 Discussion

There are a number of other human pathway resources available in the public domain. HumanCyc [9] [10] is a database with focus on metabolism, as much of it is computationally created on the basis of the EcoCyc [11] template. The data model is very similar to the Reactome data model. KEGG [5] is a curated database covering metabolic reactions and signal transduction pathways. However, different data models are used to describe these, and therefore no connections can be made between the entities involved across metabolism and signalling. Another drawback is the reliance on Enzyme Commission (EC) numbers for connecting catalysts to metabolic reactions, which can lead to ambiguous or incorrect assignments. Panther Pathways [12] is a collection of curated signaling pathways with a similar data model to Reactome. It differs, though, in its data acquisition approach, consisting of more rapid, but shallower curation. BioCarta (<http://www.biocarta.com>) and GenMAPP [13] are human pathway resources with an emphasis on data visualisation. Finally, there are the protein interaction databases like BIND [13], MINT [14] and IntAct [15] whose emphasis is on collecting high-throughput protein interaction data rather than describing the ‘mechanics’ of reactions and pathways.

When comparing the Reactome database to these other resources, it is unique in covering a wide variety of pathways found in the cell and in using a uniform data model across these pathways. This enables the user to look at proteins within an entire network of interactions rather than isolated pathways only, allowing the identification of connections that may be missed otherwise.

In conclusion, Reactome is a curated database of biological processes, describing reactions in a systematic, computationally accessible format. Reactome data are crossreferenced extensively to ensure good integration with other publically available databases. User-supplied data can be uploaded and interpreted within the Reactome reaction map. All Reactome data are freely available and can be downloaded in a variety of data formats.

References

1. Human Genome Program, U.S. Department of Energy, Genomics and Its Impact on Science and Society: A 2003 Primer (2003)
2. Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Mazumder, R., O'Donovan, C., Redaschi, N., Suzek, B.: The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.* 34 (2006) D187-D191

3. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (2000) 25-29
4. Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Hubbard, T.J.: *Ensembl. Nucl. Acids Res.* 34 (2006) D556-561
5. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. *Nucl. Acids Res.* 32 (2004) D277-280
6. Chen, F., Mackey, A.J., Stoeckert, C.J.Jr., Roos, D.S.: OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucl. Acids Res.* 34 (2006) D363-368
7. Li, L., Stoeckert, C.J. Jr., Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13 (2003) 2178-2189
8. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J.: The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models. *Bioinformatics* 19 (2003) 524-531
9. Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P., Karp, P.: MetaCyc: A multiorganism database of metabolic pathways and enzymes. *Nucl. Acids Res.* 34 (2006) D511-D516
10. Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M., Karp, P.D.: Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* 6 (2005) R2
11. Karp, P.D., Riley, M., Paley, S.M., Pelligrini-Toole, A.: EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism. *Nucl. Acids Res.* 24 (1996) 32-39
12. Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremiex, O., Campbell, M.J., Kitano, H., Thomas, P.D.: The PANTHER database of protein families, subfamilies, functions and pathways. *Nucl. Acids Res.* 33 (2005) D284-D288
13. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., Conklin, B.R.: GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.* 31 (2002) 19-20
14. Bader, G.D., Betel, D., Hogue, C.W.: BIND: the Biomolecular Interaction Network Database. *Nucl. Acids Res.* 31 (2003) 248-50
15. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G.: MINT: a Molecular INTeraction database. *FEBS Lett.* 513 (2002) 135-140
16. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R.: IntAct: an open source molecular interaction database. *Nucl. Acids Res.* 32 (2004) D452-455

Structural Similarity Mining in Semi-structured Microarray Data for Efficient Storage Construction

Jongil Jeong¹, Dongil Shin¹, Chulho Cho², and Dongkyoo Shin^{1,*}

¹Department of Computer Science and Engineering, Sejong University
98 Kunja-Dong, Kwangjin-Ku, Seoul 143-747, Korea
jijeong@gce.sejong.ac.kr, {dshin, shindk}@sejong.ac.kr

²College of Business Administration, Kyung Hee University
1 Hoegi-Dong, Dongdaemun-Ku, Seoul 130-701, Korea
rocho@unitel.co.kr

Abstract. Many researches related to storing XML data have been performed and some of them proposed methods to improve the performance of databases by reducing the joins between tables. Those methods are very efficient in deriving and optimizing tables from a DTD or XML schema in which elements and attributes are defined. Nevertheless, those methods are not effective in an XML schema for biological information such as microarray data because even though microarray data have complex hierarchies just a few core values of microarray data repeatedly appear in the hierarchies. In this paper, we propose a new algorithm to extract core features which is repeatedly occurs in an XML schema for biological information, and elucidate how to improve classification speed and efficiency by using a decision tree rather than pattern matching in classifying structural similarities. We designed a database for storing biological information using features extracted by our algorithm. By experimentation, we showed that the proposed classification algorithm also reduced the number of joins between tables.

Keywords: structural similarity, decision tree, semi-structured data, microarray database, bioinformatics, schema mining.

1 Introduction

Many researches related to storing XML data have been performed [1], [2], [3], [4]. Such researches proposed methods to improve the performance of databases by reducing the joins between tables. The proposed methods are very efficient in deriving and optimizing tables from a DTD or XML schema in which elements and attributes are defined. Nevertheless, those methods are not effective in an XML schema for biological information such as microarray data because even though microarray data have complex hierarchies just a few core values of microarray data

*Correspondence Author. This study was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea. (0412-MI01-0416-0002).

repeatedly appear in the hierarchies. Therefore, it is needed a new approach for the hierarchies of biological information.

In this paper, we present a new algorithm that optimizes biological information by extracting core features repeatedly occurs in an XML schema. In contrast with the existing researches, the new algorithm classifies elements with structural similarities within a limited scope. These features are used to form a decision tree which leads to optimal storage structure.

This paper consists of six sections. In section 2, we introduce a microarray, MAGE, and a decision tree. In section 3, we define the rules for classifying elements with structural similarities. In section 4, we extract features from an XML schema for MAGE-ML using a decision tree and then design a database optimized to biological information using these features and evaluate the performance of the database. Finally, we conclude this paper in Section 5.

2 Background

Inline technique [1], [2], [3], [4] is a rule for reducing the complexity of a DTD or an XML schema by eliminating removable elements. When an attribute is specified into a DTD or XML schema, the frequency of attribute values can be expressed by operators such as “?”, “*”, and “+”. The basic concept of inline technique is that if an element has only one sub-element in which an attribute with multiple values is defined, the attribute can be inlined into attribute of the upper level element.

In relation to this concept, research [2] proposed three order encoding methods that can be used to represent XML order in the relational data model and extended it to inline technique. Research [3], [4] proposed three types of transformations set which is similar with the rule of inline technique. Also, various researches related to schema mining have been performed. Mainly these researches proposed efficient ways to extract commonly occurring schemas in a collection of schemas and then created a common schema for query using those [5] [6]. However, such mediator schema does not help to store XML based data in efficient way.

In this section, we illustrate the decision tree algorithm, which is used as the main tool for selecting common structures from semi-structured bioinformatics data schema, and MAGE (Microarray Gene Expression) data, which is used as experimental data.

2.1 Decision Tree

A Decision Tree makes rules for classifying the patterns that exist in attribute sets and presents such rules for making decisions in a tree structure. This classification model is used to classify new records and expected values from such records. Since this technique forms the basis of classification and expectation for records, it is easy to understand classification rules, correlations, and effects between variables. Therefore, a Decision Tree makes it easy to choose target data and is a simple model [7].

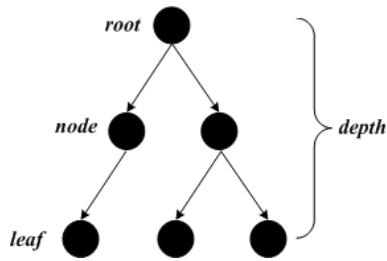


Fig. 1. The structure of a Decision Tree

In Figure 1, each black circle is a *node* or set of expected values. If a *node* is on the top level, the *node* is called the *root node*. The *root node* is a set of all expected values. A *node* that has been divided by a specific rule is called a *child node*. This procedure is called *pruning*. A *Node* at a point where *pruning* no longer occurs is called a *leaf node*. All *nodes* except *leaf nodes* are called *branch nodes*. Each path from a *root node* to a *leaf node* represents a *condition* to classify classes.

2.2 Microarray Data and MAGE (Microarray Gene Expression)

Through information generated from Microarray experiment, research groups can learn various experimental techniques used by others and make use of those to solve their biological question. To share Microarray data with other research groups, the information, which is necessary to analyze gene expression data, needs to be classified. Additionally, the classified information should be interoperable enough to exchange with other research groups. For these necessities, the Microarray Gene Expression Data (MGED) group designed an exchange model and format for The Minimal Information for the Annotation of a Microarray Experiment (MIAME) compliant data. The model and format has been announced as a standard for exchanging Microarray data: MAGE-OM (Object Model) and MAGE-ML (Markup Language) [8] [9]. All properties in MAGE-OM should be transformed to the exchange-format using MAGE-ML expressed in XML, in order to be transmitted to other research groups [10]. The bioinformatics data based on MAGE-ML is enormous, causing difficult problems for transmission and storage. This MAGE-ML data is publicly open to every researcher as a DTD (Document Type Definition) file.

3 Classification of Core Features Using Decision Tree Algorithm

In this section, we define the terminology to be used in making rules for classifying XML elements with structural similarity from semi-structured data. The defined terminology is explained as follows. Structural expression using tree nodes based on the terminology and its example XML schema are shown in Figure 2.

- e : an element defined in XML schema
- E : an elements set of e
- SE : a sub-elements set of e
- a : an attribute of e

- *A*: an attributes set of *e*
- *SA*: an attributes set for all sub-elements of *e*
- *complexType*: Structural information that consists of *SE* and (or) *A* of *e*.
- *Lowest child*: an element without a sub-element
- *Lowest parent*: an element with a sub-element that is one of the lowest child elements
- *PG (Parent Group)*: a set of candidate elements to be parents of a *Lowest Child*
- *LPCG (The Lowest Parent Candidate Group)*: a set of candidates to be *Lowest Parent*
- *LCG (The Lowest Child Group)*: a set of *Lowest child* elements
- *LPG (The Lowest Parent Group)*: a set of *Lowest Parent* elements
- *ULPG (Upper Level Parent Group)*: a set of upper level parents, including elements that are neither *Lowest Child* nor *Lowest Parent*

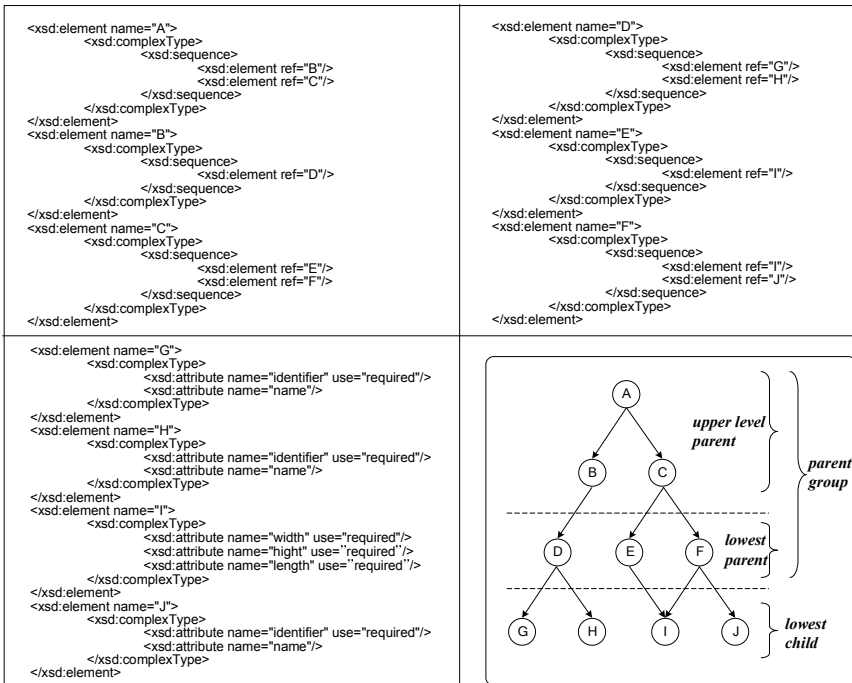


Fig. 2. Structural expression using Tree nodes and its example XML schema

As you can see in Figure 2, every element defined in an XML schema has a *complexType* that defines components, such as sub-element(s) and attribute(s). Each component can be expressed as part of a set. If a certain element *elex* has a *complexType* then each component set of *elex* is defined as follows:

$$SE_{elex} = \{e_1, e_2, \dots, e_n\}, SA_{elex} = \{A_{e1}, A_{e2}, \dots, A_{en}\}$$

$$complexType(elex) = \{SE_{elex}, SA_{elex}\}$$

When $SE_{e_{lex}}$ and (or) $SA_{e_{lex}}$ in the *complexType* of a certain element *elex* exactly match $SE_{e_{ley}}$ and (or) $SA_{e_{ley}}$ in the *complexType* of another element *eley*, we say that *elex* and *eley* have the same *complexType*.

3.1 Classification Rules

Based on the terminology and the definition of *complexType*, we define four rules for classifying similar elements which are used as branching conditions in the decision tree.

- Rule 1: If an element has no sub-elements then the element (for example, nodes G, H, I and J in Figure 1) is classified into an element of *LCG*. Otherwise, the element (for example, nodes A, B, C, D, E, and F in Figure 1) is classified into an element of *PG*. That is, Rule 1 decides that an element should belong to group *LCG* or to group *PG*. This rule is expressed by the following pseudo code.

```

For each  $e_i \in E$  {
  if(number of elements in  $SE_{e_i} == 0$ ) {
     $e_i$  is classified into LCG;
  } else {
     $e_i$  is classified into PG;
  }
}

```

- Rule 2: Let *LCG* from Rule 1 be *LCG₀*. When several elements in *LCG₀* have *complexType* in which one or more attributes are defined, the elements which have the same *complexType* are separated into a new group *LCG_p* ($p > 0$). If there is already a group *LCG_p* with the same *complexType*, the element comes to the group *LCG_p*. That is, Rule 2 classifies multiple sets of *LCG*.

```

 $p = 0$ ;
For each  $e_i \in LCG_0$  {
  Flag=0;
  If ( $p > 0$ ) {
    For  $q = 1$  to  $p$ 
      If ( $complexType(e_i) = complexType(element\ in\ LCG_q)$ ) {
         $e_i$  is classified into LCGq;
        Flag=1;
      }
    }
  If (Flag==0) {
    For each  $e_j \in LCG_0$  {
      if ( $complexType(e_i) = complexType(e_j)$ ) {
         $p = p + 1$ ;
         $e_i$  and  $e_j$  are classified into a new group of LCGp;
      }
    }
  }
}

```

- Rule 3: If a certain element in *PG* has only a sub-element that belongs to *LCG*, then the element is classified into an element of *LPG*. Otherwise, the element is classified into an element of *ULPG*. That is, Rule 3 separates elements in *PG* into two groups: *LPG* and *ULPG*.

```

For each  $e_i \in PG$  {
  if ( $SE_{e_i} \in LCG$ ) {
     $e_i$  is classified into LPG;
  } else {
     $e_i$  is classified into ULPG;
  }
}

```

- Rule 4: Let LPG from Rule 3 be LPG_0 . When several elements in LPG_0 have $complexType$ in which one or more attributes are defined, the elements which have the same $complexType$ are separated into a new group LPG_p ($p > 0$). If there is already a group LPG_p with the same $complexType$, the element comes to the group LPG_p . That is, Rule 4 classifies multiple sets of LPG .

```

p = 0;
For each  $e_i \in LPG_0$  {
  Flag=0;
  If ( $p > 0$ ) {
    For  $q=1$  to  $p$ 
      If ( $complexType(e_i) = complexType(element\ in\ LPG_q)$  {
         $e_i$  is classified into  $LPG_q$ ;
        Flag=1;
      }
    }
  If (Flag==0) {
    For each  $e_j \in LPG_0$  {
      if ( $complexType(e_i) = complexType(e_j)$  {
         $p=p+1$ ;
         $e_i$  and  $e_j$  are classified into a new group of  $LPG_p$ ;
      }
    }
  }
}

```

3.2 Decision Tree for Recognizing the Core Features

Figure 3 shows the simple graph for a decision tree made of two sub-trees. The defined decision tree has 7 branch nodes and 3 leaf nodes and the maximum level of this tree is 3. After performing rule 1, every element will be divided into two nodes: nodes without sub-elements (LCG) and nodes with sub-elements (PG).

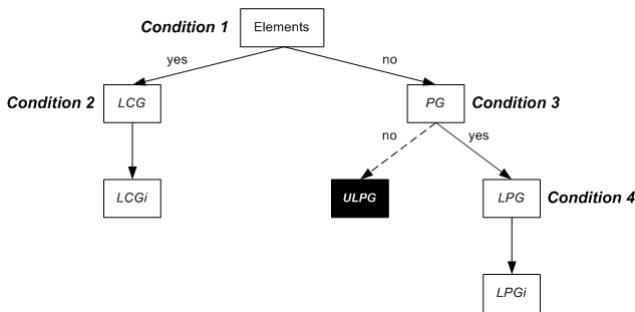


Fig. 3. The simple graph of a decision tree

Conditions 1, 2, 3, and 4 are pointing target elements to which rule, 1, 2, 3, or 4, should be applied. For an element to satisfy a rule, the $complexType$ of the element must exactly match the condition defined in the *if* statement of the rule. These conditions can be stated as follows:

- Condition 1: If rule 1 is satisfied, then e arrives at LCG . Otherwise, it arrives at PG .
- Condition 2: If rule 2 is satisfied, then e and its similar element e arrive at a new LCG .

- Condition 3: If rule 3 is satisfied, then e arrives at LPG . Otherwise, it arrives at $ULPG$.

Condition 4: If rule 4 is satisfied, then e and elements similar to e arrive at a new LPG .

4 Database Design Using the Proposed Decision Tree Algorithm

We design a database table from the classified XML schema for MAGE-ML data using the decision tree illustrated in the previous section.

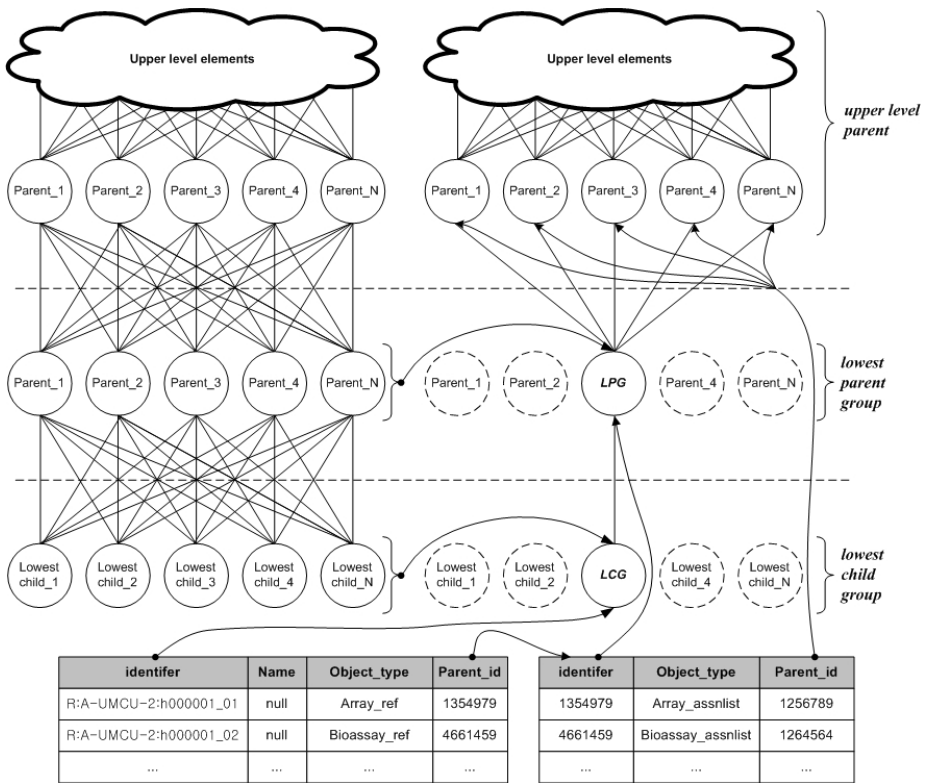


Fig. 4. Classified and integrated database tables

Figure 4 shows database tables created for LCG and LPG . Although several tables for multiple sets of LCG and LPG can be made respectively, we have just two tables, because the XML schema for MAGE-ML has just one group each for LCG and LPG respectively. (The case of MEGE-ML is suitable for illustration purpose.) The two tables for LCG and LPG have *identifiers* to ensure the uniqueness of a row and *object_type*, which denotes the Object type of a row. By having an identifier, it is possible to maintain the hierarchy between the lowest parent and the lowest child. As

shown in Figure 4, there is a *one to one join* between *LCG* and *LPG* but there are *one to N* joins between *LPG* and *ULPG*. This means that the *N to N* joins between *LCG* and *LPG* for the raw XML data are reduced to one to *N* joins for the classified XML data. This leads to an improvement in the complexity of the database. Since *LCG* and *LPG* are the prominent data and occur repeatedly, we focused on a lowest child group (*LCG*) and a lowest parent group (*LPG*) in the classifying rules.

4.1 Time Complexity

From the improvement in the complexity of the database space, the performance in storing and loading XML data is improved. The XML data used in the experiment is a document of 1Mbyte size that has all its elements classified into *LCG* and *LPG*.

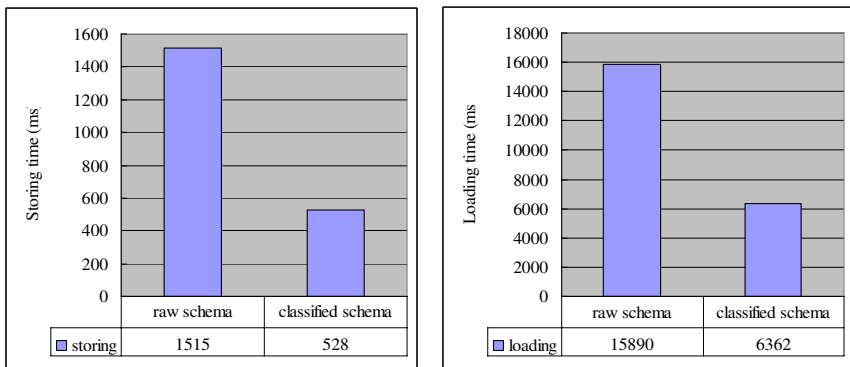


Fig. 5. The comparison of storing time and loading time

As shown in Figure 5, the proposed scheme saves processing time for storing and loading XML data. This means that the reduction of table joins plays an important role in saving time in storing and loading XML data [3].

4.2 Database Space Complexity

We achieved an enhancement on the complexity of a database storing XML data by applying the decision tree scheme to MAGE-ML as shown in Table 1. Since we created tables according to class rules, in which a class is mapped to a table, the total number of classes and tables are the same.

Table 1. Database space complexity

	Raw schema	Classified schema
Total classes	455	314
Total tables	455	314
Total records	2012	160
Total DB size	710 (Kb)	27 (Kb)
Total table joins	101	2

4.3 Reconstructing the XML Document

The existing research [2] accesses a set of tables to retrieve tuples and unions these results together and later orders them to form XML Document using the efficient sorted outer-union method. In order to simplify this task, we used a tool for automatically generating the XML Document from relational databases.

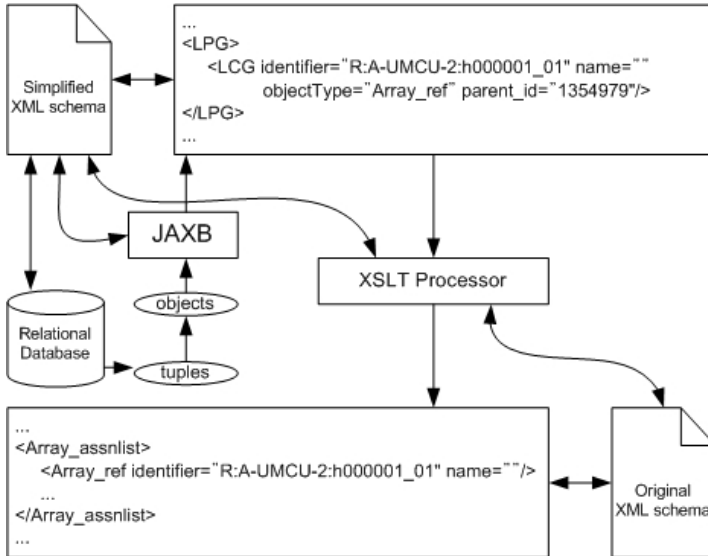


Fig. 6. The flow of reconstructing the XML Document

JAXB (Java Architecture for XML Binding) [11] provides the functionality to convert an object to an XML Document and vice versa. We retrieve tuples from relational database and map the results to objects, and then convert those to XML Document using JAXB. Thanks to the XML schema centric database design, it is very efficient to generate the XML Document from relational database via JAXB. However, this XML Document is not complete because the XML schema was already simplified before performing the XML schema centric database design. We used XSLT (XML Stylesheet Language Transformations) [12] to reconstruct the complete XML Document from it. As you can see in Figure 4, each of the lowest child elements has not only object_type but also parent_id, hence, XSLT processor can easily aware the order specified in the original XML schema. Figure 6 shows the flow we implemented.

5 Conclusion

This paper proposed a scheme for mining similar elements with structural similarities from an XML schema for biological information, in which a number of elements and attributes are defined. The experimental results show that our scheme improves the

performance of storing and loading XML data by reducing the joins between tables remarkably. Therefore, the scheme proposed in this paper presents a way to optimize a database when designing XML schema centric databases.

In future work, we plan to extend the rules for classifying similar elements. In the proposed method, we limited the classification scope to *LCG* to *LPG* so as to minimize the transformation of the hierarchy of the XML schema. If we widen the scope from *LCG* to *ULPG*, database tables and joins between tables will be reduced so that the performance for storing and loading XML data will be improved over the current result.

References

1. Schoning, H.: Tamino - A DBMS designed for XML, In Proceedings of the 17th ICDE Conference, Heidelberg, Germany (2001) 149-154
2. Tatarinov, I., and Viglas, S. D.: Storing and Querying Ordered XML Using a Relational Database System, Proceedings of the 2002 ACM SIGMODACM SIGMOD international conference on Management of data, Madison, Wisconsin (2002) 204-215
3. Shanmugasundaram, J., Tufte, K., He, G., Zhang, C., DeWitz, D., and Naughton, J.: Relational databases for querying xml documents: Limitations and opportunities, In Proc. Intl. Conf. on 25th VLDB, 1999
4. Runapongsa, K., and Patel, J. M.: Storing and Querying XML Data in Object-Relational DBMSs, EDBT Workshop XMLDM 2002 266-285
5. Laur, P.A., Masegla, F., and Poncelet, P.: Schema Mining: Finding Structural Regularity among Semistructured Data, Lecture Notes in Computer Science, Volume 1910 (2000) 498
6. Anne, L., Pascal, P., and Maguelonne, T.: Towards a fuzzy approach for mining XML mediator schemas, Fuzzy Logic and the Semantic Web Workshop, 2005
7. Witten, I. H., and Frank, E.: Data Mining Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers
8. Wang, H., Li, J., Luo, J., and He, Z.: XCPaqs: Compression of XML Document with XPath Query Support, 2004 IEEE, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), 2004
9. Sarkans U, Parkinson H, Lara GG, Oezcimen A, Sharma A, Abeygunawardena N, Contrino S, Holloway E, Rocca-Serra P, Mukherjee G, Shojatalab M, Kapushesky M, Sansone SA, Farne A, Rayner T, Brazma A.: The ArrayExpress gene expression database: a software engineering and implementation perspective, Bioinformatics 2005, 21: 1495-1501.
10. Levene, M., and Wood, P.: XML Structure Compression, In Proc. 2nd Int. Workshop on Web Dynamics, 2002
11. JAXB (Java Architecture for XML Binding): <http://java.sun.com/xml/downloads/jaxb.html>
12. XSLT (XML Stylesheet Language Transformations): <http://www.w3.org/Style/XSL/>

Modeling and Storing Scientific Protocols

Natalia Kwasnikowska¹, Yi Chen², and Zoé Lacroix²

¹ Hasselt University and Transnational University of Limburg, Belgium

`natalia.kwasnikowska@uhasselt.be`

² Arizona State University, Tempe, AZ, USA

`{yi, zoe.lacroix}@asu.edu`

`http://bioinformatics.eas.asu.edu`

Abstract. We propose an abstract model for scientific protocols, where several atomic operators are proposed for protocol composition. We distinguish two different layers associated with scientific protocols: design and implementation, and discuss the mapping between them. We illustrate our approach with a representative example and describe ProtocolDB, a scientific protocol repository currently in development. Our approach benefits scientists by allowing the archiving of scientific protocols with the collected data sets to constitute a scientific portfolio for the laboratory to query, compare and revise protocols.

1 Introduction

Scientific discovery relies on the adequate expression, execution, and analysis of scientific protocols. Although data sets are properly stored, the protocols themselves are often recorded only on paper or remain in a digital form developed to implement them. Once the scientist who has implemented the scientific protocol leaves the laboratory, the record of the scientific protocol may be lost. Collected data sets without the description of the process that produced them may become meaningless. Moreover, to support scientific discovery, anyone should be able to reproduce the experiment. Therefore, a detailed description of the protocol is necessary, together with the collected data sets.

A scientific protocol is the process that describes the experimental component of scientific reasoning. Scientific reasoning follows a hypothetico-deductive pattern and is composed of the succession of the expression of a *causal question*, a *hypothesis*, the *predicted results*, the design of an *experiment*, the actual *results* of the experiment, the comparison of the predicted and the experimental results, and the *conclusion*, supportive or not of the hypothesis [1]. Scientific protocols (also called data-analysis pipelines, workflows or dataflows) are complex procedural processes composed of a succession of tasks expressing the way the experiment is conducted. They usually involve a data-gathering stage, that may be followed by an analysis stage. A scientific protocol thus describes how the experiment is conducted and records all necessary information to reproduce the experiment. In bioinformatics, the importance of identifying protocol tasks has been addressed by Stevens et al. [2] and Bartlett et al [3], while Tröger [4]

has proposed a language for expressing *in silico* protocols that approximates research method used for *in vitro* experiments.

We propose a high-level abstract model for scientific protocols representing two different layers associated with scientific protocols: design and implementation, and discuss the mapping between them. Our approach benefits scientists by allowing the archiving of scientific protocols with data sets to constitute a scientific portfolio for the laboratory to query, compare and revise protocols.

2 Related Work

Several approaches integrate scientific protocol models with a database system, but provide little support for the actual design phase of a protocol, do not distinguish between design and implementation, and provide limited support for querying and versioning of protocols. They include the Object Protocol Model [5] and Zoo [6] that both use the object-oriented data model. More recent efforts propose an integration of protocols and relational databases. Shankar et al. [7] propose a language for modeling protocols that is tightly integrated with SQL.

On the other hand, several systems focus on the design issues of protocols, sometimes combined with the (distributed) execution of protocols, but without fully leveraging the storage and query capability of databases. They include Taverna [8], with a vast integration of bioinformatics resources, and Kepler [9,10], based on the Ptolemy II system. WOODSS [11] emphasizes the support of several abstraction levels of protocol design and facilitates protocol composition and reuse. Several researchers have agreed on the separation of the design of a protocol from its implementation [10,12,13,14]. For instance, Ludäscher et al. [10] propose a distinction between abstract and concrete protocols and use database mediation techniques for an abstract-to-concrete translation. Zhao et al. [14] propose an XML-based virtual data language for a typed and compositional protocol specification, with mapping descriptors between the design and implementation. A formal graphical language for hierarchical modeling of protocols has also been proposed in Hidders et al. [15], and combines Petri nets with operators and typing system from nested relational calculus.

Compared with previous work, we focus here on how to define a formal, abstract model for defining scientific protocols that is as high-level as possible, so as to be suitable to general applications as well as for storage of protocols in a database system. We distinguish the design from possible implementations of protocols and define the mapping between them.

3 Modeling Scientific Protocols

The abstract protocol definition language introduced in this section aims at representing the structure of scientific protocols, and is, by design, unbiased towards any specific data model or query language. We aim at modeling *in vivo*, *in vitro*, as well as *in silico* experiments.

Each step of a scientific protocol can be represented by a *task* [2,3]. To model a protocol, we distinguish its *design*, that captures its scientific aim, from its *implementation*, that specifies resources selected to execute the tasks. This distinction allows for comparison between different choices of resources, allowing the scientist to select the implementation best meeting the protocol's needs. Therefore we decompose each scientific protocol into two components: *protocol design* and *protocol implementation*. Both components consist of coordinated *tasks*, but at different abstraction levels. As we present a syntactical model, we will specify the data flow by identifying its *conceptual type* and *format*.

Each task of the protocol design is defined by its *task name*, *conceptual input type*, and *conceptual output type*. When an ontology is available to describe the scientific objects and tasks involved, the input and output of each protocol design task may be defined by their respective concept classes. The protocol design task itself may appear in the ontology, as a relationship defined between the input concept class and output concept class.

A task of the protocol implementation describes the resource selected to implement a protocol design task. Each protocol implementation task is defined by its *application name*, *input format*, and *output format*. The input/output format is a possible representation for the conceptual input/output type of the corresponding design task. The name of a protocol implementation task denotes a resource, an application or a service, implementing the corresponding protocol design task, together with its annotation (e.g., parameters, url etc).

3.1 Protocol Design Model

A scientific protocol can be defined inductively from tasks, or basic protocols, and four connectors. Formally, the protocol design model is defined as follows.

Definition 1. Let \mathcal{T} be a set of task names. Let \mathcal{C} be a set of conceptual type names, over which an operator \oplus is defined and a sub-typing relation " \preceq ". A protocol design task \mathbb{T}_D is a triple (i, n, o) with $i, o \in \mathcal{C}$ and $n \in \mathcal{T}$. The set $\mathbb{T}_{\mathcal{T}, \mathcal{C}} = \mathcal{C} \times \mathcal{T} \times \mathcal{C}$ is the set of protocol design tasks defined from \mathcal{T} and \mathcal{C} . Now we define recursively the set $\mathbb{P}_{\mathcal{T}, \mathcal{C}}$ of protocol designs defined from \mathcal{T} and \mathcal{C} , and for each protocol design D we impose requirements on its input type $\text{In}(D)$ and its output type $\text{Out}(D)$:

- if $D = (i, n, o)$ then $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, with $\text{In}(D) = i$ and $\text{Out}(D) = o$;
- if $D_1 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $\text{Out}(D_1) \preceq \text{In}(D_2)$ then $D_1 \cdot D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, with $\text{In}(D_1 \cdot D_2) \preceq \text{In}(D_1)$ and $\text{Out}(D_2) \preceq \text{Out}(D_1 \cdot D_2)$;
- if $D_1 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ then $D_1 \oplus D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, with $\text{In}(D_1 \oplus D_2) \preceq \text{In}(D_1) \oplus \text{In}(D_2)$ and $\text{Out}(D_1) \oplus \text{Out}(D_2) \preceq \text{Out}(D_1 \oplus D_2)$;
- if $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $\text{Out}(D) \preceq \text{In}(D)$ and k is an integer then $D^k \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, with $\text{In}(D^k) \preceq \text{In}(D)$ and $\text{Out}(D) \preceq \text{Out}(D^k)$;
- if $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $\text{Out}(D) \preceq \text{In}(D)$ then $D^* \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, with $\text{In}(D^*) \preceq \text{In}(D)$ and $\text{Out}(D) \preceq \text{Out}(D^*)$.

In the above definition, the operator " \cdot " denotes the *successor* connector, i.e., the serial composition. The operator " \oplus " denotes the *split-merge* connector,

i.e., the parallel composition. The operators k and $*$ denote k -recursion, and $star$ -recursion, respectively. Relation “ \preceq ” denotes sub-typing between conceptual type names, provided by chosen ontology or type system. Type $i_1 \oplus i_2$, with $i_1, i_2 \in \mathcal{C}$, denotes a collection type, whose precise semantics depends on the semantics of the split-merge connector.

We emphasize that Def. 1 only captures the syntax of a protocol design, where the data flow is only described in terms of conceptual type names. Nevertheless, based on the interaction with our collaborators, we claim that our definition of protocol design is sufficient to faithfully model scientific protocols used in practice, once suitable semantics are provided for the operators.

Definition 2. Let $(i, n, o) \in \mathcal{C} \times \mathcal{T} \times \mathcal{C}$. We define recursively the set of types $Types(D)$ and the set of Tasks(D) of a protocol design D as follows:

- if $D = (i, n, o)$ then $Types(D) = \{i, o\}$ and $Tasks(D) = \{(i, n, o)\}$,
- if $D = D_1 \cdot D_2$ or $D = D_1 \oplus D_2$, then $Types(D) = Types(D_1) \cup Types(D_2)$ and $Tasks(D) = Tasks(D_1) \cup Tasks(D_2)$,
- if $D = D_1^k$ or $D = D_1^*$, then $Types(D) = Types(D_1)$ and $Tasks(D) = Tasks(D_1)$.

We say that a protocol design D is *composed of* the tasks in $Tasks(D)$. If a protocol design D is of the form $D_1 \cdot D_2$ or $D_1 \oplus D_2$, with $D_1, D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$, then D is *directly composed of* D_1 and D_2 . Similarly, if a protocol design D is of the form D_1^k or D_1^* , then D is *directly composed of* D_1 . If a protocol design D is composed of D_1 , then we also call D_1 a *sub-protocol* of D .

3.2 Protocol Implementation Model

Once the design of a protocol is defined, its specification in terms of resources used to execute it may be defined. Each design step may be implemented by specifying the input format, an application name, and the corresponding output format. Although sometimes a design step can be implemented by a single implementation step, it is common that a design step needs to be mapped to a complex process, involving multiple biological resources, thus to a sub-protocol rather than a single task of the implementation protocol. The need for a sub-protocol to implement a single design task may occur to include adapters to translate the output format from the previous implementation step into the input format of the selected implementation resource, or for specifying alternative implementations.

The protocol implementation model is similar to the protocol design model. Specifically, rather than a set of task names \mathcal{T} , we now have a set of application names \mathcal{A} . Rather than a set of conceptual type names \mathcal{C} , we now have a set of format names \mathcal{F} , over which an operator \oplus is defined. The set of protocol implementation tasks $\mathbb{T}_{\mathcal{A}, \mathcal{F}}$ and the set of protocol implementations $\mathbb{P}_{\mathcal{A}, \mathcal{F}}$ are defined similar to Def. 1, except that for the sake of concreteness, we replace sub-typing “ \preceq ” on conceptual type names by equality “ $=$ ” on format names.

The set $Formats(I)$ of format names and the set $Resources(I)$ of application names of an protocol implementation I , i.e., $I \in \mathbb{P}_{\mathcal{A}, \mathcal{F}}$, have a definition similar to Def. 2. It is worth noting that $Formats(I)$ and $Resources(I)$ provide basic provenance information for the data collected by executing protocol I .

3.3 Mapping Design to Implementation

Each design task of the design protocol may be mapped to one or more implementation tasks or protocols.

Definition 3. A conceptual type mapping is a partial function $\varphi_C: \mathcal{C} \rightarrow \mathcal{F}$. A protocol design task mapping is a partial function $\varphi_T: \mathbb{T}_{\mathcal{T}, \mathcal{C}} \rightarrow \mathbb{P}_{\mathcal{A}, \mathcal{F}}$. A protocol design task mapping φ_T is said to be consistent with a conceptual type mapping φ_C if for every protocol design task $T_D \in \mathbb{T}_{\mathcal{T}, \mathcal{C}}$ it holds that if $\varphi_T(T_D) = I$ then $\text{In}(I) = \varphi_C(\text{In}(T_D))$ and $\text{Out}(I) = \varphi_C(\text{Out}(T_D))$. If $\varphi_T(T_D) = I$ then we call I an implementation of protocol design task T_D under φ_T .

Definition 4. Let $D, D_1, D_2 \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$. Given a protocol design task mapping φ_T we define its generalization $\hat{\varphi}_T: \mathbb{P}_{\mathcal{T}, \mathcal{C}} \rightarrow \mathbb{P}_{\mathcal{A}, \mathcal{F}}$ such that $\hat{\varphi}_T$ corresponds to φ_T on $\mathbb{T}_{\mathcal{T}, \mathcal{C}}$ and:

- $\hat{\varphi}_T(D_1 \cdot D_2) = \hat{\varphi}_T(D_1) \cdot \hat{\varphi}_T(D_2)$,
- $\hat{\varphi}_T(D_1 \oplus D_2) = \hat{\varphi}_T(D_1) \oplus \hat{\varphi}_T(D_2)$,
- $\hat{\varphi}_T(D^k) = \hat{\varphi}_T(D)^k$ and
- $\hat{\varphi}_T(D^*) = \hat{\varphi}_T(D)^*$.

We call $\hat{\varphi}_T$ a protocol design mapping. If $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $I = \hat{\varphi}_T(D)$ then I is an implementation of protocol design D under $\hat{\varphi}_T$.

Definition 5. Let $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$. We define the set $\Phi(D)$ of all possible implementations of D and its associated mappings as the set of all tuples $(\varphi_C, \varphi_T, I)$ where:

- φ_C is a conceptual type mapping with $\text{dom}(\varphi_C) = \text{Types}(D)$,
- φ_T is a protocol design task mapping with $\text{dom}(\varphi_T) = \text{Tasks}(D)$ and consistent with the conceptual type mapping φ_C ,
- I is a protocol implementation of D under $\hat{\varphi}_T$.

Finally we define the protocol itself, composed of a protocol design and a set of protocol implementations.

Definition 6. We define a protocol $P = (D, \text{Imp}(D))$ as a pair of protocol design $D \in \mathbb{P}_{\mathcal{T}, \mathcal{C}}$ and $\text{Imp}(D)$ being a finite subset of $\Phi(D)$.

4 Example of a Scientific Protocol

We present a representative example of scientific protocol: *Study of germination of Lesquerella seeds*.¹ Lesquerella species are a promising oil crop with potential

¹ This protocol was collected at the U.S. Arid-Land Agricultural Research Center, Maricopa, AZ, courtesy of Jeff White and Neal Adam, the author of the protocol.

for industrial applications. Different members of that species possess different traits with regard to oil content, oil quality and yield. Current breeding programs aim to produce a variety suitable for commercial cultivation. One of the prerequisites to achieve this aim, is prolonged storage of seeds. The following protocol (restricted here to the *in silico* part for space reasons) was developed to determine base and optimal temperatures for germination of different Lesquerella seeds.

4.1 Statistical Analysis of Lesquerella Germination Data

The data obtained from the *in vitro* part of the experiment and stored in `Observations.xls`, was analyzed using SAS programs.

1. Determination of maximum germination was performed by succession of two SAS programs: `max` and `first obs.sas` used `Observations.xls` as input and produced `max percentages.xls` as output. That file was used as input to `merge maxmin.sas`, which produced the file `maxmin germshoots.xls`.
2. Germination proportions were analyzed using program `genmod.sas`, with the file `maxmin germshoots.xls` as input and two files as output: `maxshoots diffs.xls` and `maxshoots lsmeans.xls`.
3. Preprocessing of data necessary for determination of base and optimal temperatures for germination was achieved in two sub-steps. `Observations.xls` was used as input to `sample numbers` for `DAPest.sas` resulting in file `DAPest sample numbers.xls`, and was subsequently used as input to `DAPest.sas` which produced file `DAPestData.xls`. Also, `Observations.xls` was used as input to `graphing to print.sas`, which produced five bitmaps.
4. Base temperature (TB) for germination was determined by two separate methods, but only one of the methods was suitable for determining optimal temperatures (TO).
 - (a) TB by regression analysis — `reg.sas` was run 4 times with `DAPestData.xls` as input and producing an Excel file each time. Those four files were subsequently merged by `merge datasets.sas` into a single Excel file. That file was analyzed with `proc mixed.sas` producing two output files `TbG50mns.xls` and `TbG50mndiffs.xls`.
 - (b) TB and TO by 2-phase linear regression, broken model — the following analysis was repeated 13 times, for each kind of seed. `DAPestData.xls` was used as input to `pho341.sas`, producing an intermediate file. That file served as input to `pho342.sas`, producing another intermediate file which was used as input to `broken3.sas`. The latter produced two Excel files: `SeedID (limits).xls` and `SeedID (limits2).xls`.

4.2 Analysis of the Structural Features

Analyzing scientific protocols, we frequently observe that a single protocol step includes multiple tasks. Step 1 for determining maximum germination of seeds includes two sub-steps, each consisting of the execution of a SAS program.

The enumeration of steps does not always reflect the order of tasks. In step 4, step 4a for computing base temperature and step 4b for computing base and optimal temperatures, can be executed in parallel, although they are stated in sequential order in the example. Some steps introduce a loop, e.g., step 4b is performed for every kind of seed. This is a particular kind of loop, that can be expressed by an iteration over the “collection” of seeds.

We see that the main structure of the protocol is mostly linear (step 1 and 2), or parallel (step 4a and 4b) or introduces a loop (step 4b). We also observe that the description of the protocol mixes the design with implementation. The implementation itself can be diverse. Most steps are implemented by using applications, but sometimes manual interaction may be necessary.

4.3 Example Protocol Model

The protocol presented in Sect. 4.1 can be modeled with the definitions of Sect. 3 as follows. First, we define the set of type names \mathcal{C} as $\{\mathbf{SeedData}\}$ and the set of design task names \mathcal{T} as $\{\text{MaxGermination, Proportions, Preprocessing, BaseTemp, BaseAndOptTemp}\}$. We define the following protocol design tasks:

- T_{D1} : (**SeedData**, MaxGermination, **SeedData**),
- T_{D2} : (**SeedData**, Proportions, **SeedData**),
- T_{D3} : (**SeedData**, Preprocessing, **SeedData**),
- T_{D4} : (**SeedData**, BaseTemp, **SeedData**),
- T_{D5} : (**SeedData**, BaseAndOptTemp, **SeedData**).

We define now protocol design D with input $\text{In}(D) = \mathbf{SeedData}$ and output $\text{Out}(D) = \mathbf{SeedData}$ as $D = D_1 \oplus D_2$, with $D_1 = T_{D1} \cdot T_{D2}$, $D_2 = T_{D3} \cdot D_3$, $D_3 = T_{D4} \oplus D_4$ and $D_4 = T_{D5}$ ¹³. Note that D_1 corresponds to steps 1 and 2 in our example, D_2 to steps 3 and 4, D_3 to step 4 and D_4 to step 4b. Last but not least, D represents the design of the whole protocol (left-hand side of Fig. 1).

Because the description presented in Sect. 4.1 is an actual implementation, our protocol implementation follows it closely, with following simplifications. Whenever multiple Excel files were generated, we assume they could have been equally merged into one file with multiple tabs. If the multiple outputs have different format names, additional converters are introduced. We plan to address the issue of multiple outputs in the future when we define operator semantics.

We define the set of format names \mathcal{F} as $\{\text{Excel, Bitmap}\}$ and we simply use the set of program names as \mathcal{A} . The protocol implementation tasks are:

- T_{I1} : (Excel, max and first obs.sas, Excel),
- T_{I2} : (Excel, merge maxmin.sas, Excel),
- T_{I3} : (Excel, genmod.sas, Excel),
- T_{I4} : (Excel, sample numbers for DAPest.sas, Excel),
- T_{I5} : (Excel, DAPest.sas, Excel),
- T_{I6} : (Excel, graphing to print.sas, Bitmap),
- $T_{I6'}$: (Bitmap, convert2excel.exe, Excel),
- T_{I7} : (Excel, reg.sas, Excel),
- T_{I8} : (Excel, merge datasets.sas, Excel),
- T_{I9} : (Excel, proc mixed.sas, Excel),

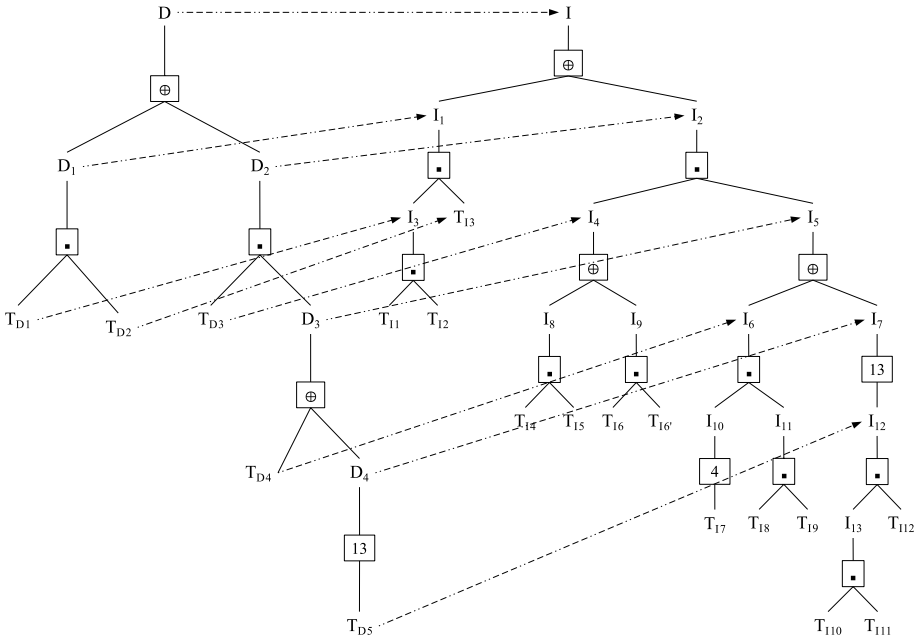


Fig. 1. Design (left) and implementation (right) of example presented in Sect. 4.1

- T_{I10} : (Excel, pho341.sas, Excel),
- T_{I11} : (Excel, pho342.sas, Excel),
- T_{I12} : (Excel, broken3.sas, Excel).

We define now protocol implementation I with input $In(I) = \text{Excel}$ and output $Out(I) = \text{Excel}$ as $I = I_1 \oplus I_2$, with $I_1 = I_3 \cdot T_{I3}$, $I_2 = I_4 \cdot I_5$, $I_3 = T_{I11} \cdot T_{I12}$, $I_4 = I_8 \oplus I_9$, $I_8 = T_{I14} \cdot T_{I15}$, $I_9 = T_{I16} \cdot T_{I16'}$, $I_6 = I_{10} \cdot I_{11}$, $I_{10} = T_{I17}^4$, $I_{11} = T_{I18} \cdot T_{I19}$, $I_7 = I_{12}^{13}$, $I_{12} = I_{13} \cdot T_{I112}$ and $I_{13} = T_{I110} \cdot T_{I111}$. The protocol implementation is illustrated in the right-hand side of Fig. 1.

Now we are ready to define the mapping between protocol design D and protocol implementation I . We define φ_C as a mapping on \mathcal{C} with $\varphi_C(\text{SeedData}) = \text{Excel}$. The protocol design mapping is represented in the picture by dashed lines. Finally, the whole protocol is defined as $(D, \{\varphi_C, \varphi_T\}, I)$.

5 ProtocolDB

ProtocolDB is composed of an Access database as the back-end repository on the server-side. The server side components also include the Microsoft Internet Information Server (Version 5) and the Apache Tomcat Web Server (Version 5.5.16) to handle user requests. The user interface interacts with these sub-systems to provide the necessary storage and retrieval functionalities. HTTP connectors

and JDBC-ODBC connectivity functionality are used to communicate with the database to perform the operations on the repository.

The schema is composed of five tables. The table `registered_users` stores the information related to the users of the system. The primary key for this table is the attribute `nickname`. Once a scientist is registered, each entry or modification of a protocol will be recorded with the scientist's information. Information pertaining to the protocol as a whole, including its name, scientific aim, date and time of last edition are stored in `protocol_info`. The attributes `DateSaved`, `TimeSaved` and `ProtocolNickname` form the composite primary key for this relation. Design steps of protocols are stored in the table `Design_Steps`. The attributes `StepNumber`, `DateSaved`, `TimeSaved`, and `ProtocolNickname` form the composite primary key for this relation. Similarly, implementation steps are stored in the table `Implementation_Steps`. The structure of the protocol consisting of the successor connector “.” and the split-merge connector “ \oplus ” is recorded in table `connection_between_steps`. The attributes `from_step`, `to_step`, `ProtocolNickname`, `DateSaved`, and `TimeSaved` form the composite primary key for this relation. The foreign key of each of the relation `Design_Steps`, `Implementation_Steps`, and `connection _between_steps` is linked to the primary key (composite primary key) of the `protocol_info` relation to link the information corresponding to a particular protocol distributed among the different relations in the database.

6 Conclusion

The model for scientific protocols proposed in the paper aims at representing the general structure of scientific protocols, with explicit distinction between design and implementation. It allows for comparison of alternative implementations and resources, an approach that is compatible with path-based guiding systems [16]. On-going and future work includes the extension of the model to allow the storage of collected data sets, for support of cross protocol-data queries and reasoning on data provenance.

This model is used to develop ProtocolDB, a repository of scientific protocols where scientists can store, retrieve, compare, and re-use scientific protocols. The system is currently under development and will be available shortly at:

<http://bioinformatics.eas.asu.edu/protocoleDatabase.htm>.

Acknowledgments. This research was partially supported by the National Science Foundation grants IIS 0612273, IIS 02230042 and IIS 0222847, and the NIH National Library of Medicine grant R03 LM008046-01. ProtocolDB was implemented by Phanindra Dev Deepthimahanthi. We thank Jeff White, U.S. Arid-Land Agricultural Research Center (ALARC), Mike Berens, Anna Joy, and Dominique Hoelzinger, Translational Genomics Research Institute, Jan Van den Bussche, Hasselt University, and Jan Hidders, University of Antwerp, for their valuable input.

References

1. Lawson, A.: *Studying for Biology*. Addison-Wesley Educational Publishers (1995)
2. Stevens, R., Goble, C.A., Baker, P.G., Brass, A.: A classification of tasks in bioinformatics. *Bioinformatics* **17**(1) (2001) 180–188
3. Bartlett, J.C., Toms, E.G.: Developing a Protocol for Bioinformatics Analysis: An Integrated Information Behavior and Task Analysis Approach. *Journal of the American Society for Information Science and Technology* **56**(5) (2005) 469–482
4. Tröger, A., Fernandes, A.: A language for comprehensively supporting the in vitro experimental process in silico. In: Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2004). (2004) 47–56
5. Chen, I.M.A., Markowitz, V.M.: An overview of the object protocol model (opm) and the opm data management tools. *Inf. Syst.* **20**(5) (1995) 393–418
6. Ailamaki, A., Ioannidis, Y.E., Livny, M.: Scientific workflow management by database management. In: SSDBM. (1998) 190–199
7. Shankar, S., Kini, A., DeWitt, D.J., Naughton, J.: Integrating databases and workflow systems. *SIGMOD Rec.* **34**(3) (2005) 5–11
8. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., Li, P.: Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics Journal* **20**(17) (2004) 3045–3054
9. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J., Zhao, Y.: Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows* (2005) (to appear).
10. Ludäscher, B., Altintas, I., Gupta, A.: Compiling abstract scientific workflows into web service workflows. In: SSDBM, IEEE Computer Society (2003) 251–254
11. Medeiros, C., Alcazar, J., Digiampietri, L., Pastorello, G., Santanche, A., Torres, R., Madeira, E., E, B.: WOODSS and the Web: annotating and reusing scientific workflows. *SIGMOD Record* **34**(3) (2005) 18–23
12. Hashmi, N., Lee, S., Cummings, M.: Abstracting Workflows: Unifying Bioinformatics Task Conceptualization and Specification Through Semantic Web Services. In: *W3C Semantic Web for Life Sciences position paper*. (2004)
13. Foster, I., Voekler, J., M.Wilde, Y.Zhao: Chimera: a virtual data system for representing, querying and automating data derivation. In: SSDBM. (2002) 37
14. Zhao, Y., Dobson, J., Foster, I., Moreau, L., Wilde, M.: A notation and system for expressing and executing cleanly typed workflows on messy scientific data. *ACM SIGMOD Record* **34** (2005) 37–43
15. Hidders, J., Kwasnikowska, N., Sroka, J., Tyszkiewicz, J., Van den Bussche, J.: Petri net + nested relational calculus = dataflow. Technical Report TR UA 2006-04, University of Antwerp, Belgium (2006)
16. Cohen-Boulakia, S., Davidson, S., Froidevaux, C., Lacroix, Z., Vidal, M.E.: Path-based systems to guide scientists in the maze of biological data sources. *Journal of Bioinformatics and Computational Biology* (2006) (to appear).

A Knuckles-and-Nodes Approach to the Integration of Microbiological Resource Data

Bart Van Brabant¹, Peter Dawyndt^{1,2}, Bernard De Baets², and Paul De Vos¹

¹ Laboratory of Microbiology

K.L. Ledeganckstraat 35, B-9000 Ghent, Belgium, Ghent University

² Department of Applied Mathematics, Biometrics and Process Control
Coupure links 653, B-9000 Ghent, Belgium, Ghent University

Abstract. Providing elementary resources to research institutes, storing a part of the world's biodiversity and carrying out essential research, biological resource centers (BRCs) fulfil an important role in contemporary life sciences. There are currently over 500 BRCs registered at the World Data Center for Microorganisms. All of these institutes offer information on their cultures independently and there is no efficient linkage between the biological material and its related data held by information providers such as Genbank, Swissprot and PubMed. As a result, researchers searching for information on microbial strains or species frequently encounter inconsistencies. This paper presents the StrainInfo.net bioportal for integrating the information provided by BRCs and third party information holders. It is shown how intelligent data integration is the best protection against information overkill and is a necessary precursor for more advanced data mining to further exploit the assets of BRCs. To establish a solid service to the research community, the suggested integration is accomplished following the *knuckles-and-nodes* approach. Initially implemented as a web interface, the underlying system also provides possibilities for setting up web services. The StrainInfo.net bioportal will be made available at <http://www.StrainInfo.net>.

1 Introduction

Once a purely experimental and observational activity, hardly involving complex data flows, biology nowadays is a highly data-dependent and computation intensive science. With the coming of age of the Internet, the scientific community attempts to improve access to the information it extracts from its research efforts. However, with its terabits of 'very dirty' data and numerous subdomains each using different terminologies and concepts, biology now faces the complex problem of data integration. With huge amounts of data scattered over a multitude of private and public databases, integration inevitably must cope with problems like data duplication, inconsistent data, synchronisation and intellectual property right issues. The relatively new field of bioinformatics uses concepts and techniques stemming from several scientific areas such as computer science, mathematics and statistics to tackle some of these hurdles. The StrainInfo.net

bioportal offers a structured and integrated view on microbial resource information. In this paper we will describe how complex information networks such as the one provided by the StrainInfo.net bioportal can be used to detect and resolve deeply hidden data inconsistencies, as well as to derive new information relationships that directly result from the merger of disparate information sources.

2 Integration Architecture

2.1 Biological Resource Centers

Biological Resource Centers (BRCs) play a major role in life sciences and biotechnological innovation. They provide researchers and companies with the specimen they need for experimentation and R&D, and store these resources in agreement with quality and biosafety regulations. Most of these BRCs maintain an online catalogue facilitating access to the information on their assets. A single microbial resource is referred to as a culture, a specific instance of the strain it belongs to. Strains generally originate from microbial isolates that are cultured and spread among different BRCs by subculturing [1].

Multiple BRCs thus each hold a different culture of the same strain. This may bring along certain ambiguities, as the institutions use different naming strategies, giving rise to different species names (synonyms) assigned to cultures of the same biological material. Moreover, collections sometimes classify a single strain in different taxonomic subgroups or give resembling names to cultures of different strains. These properties obviously make efficient data integration very hard. Although the CABRI project [2] provides an initial framework to standardize the data exchange formats used, many hurdles still need to be taken: synchronization between a federated system and its data providers, information linkage between BRCs and third party information providers (sequence data, published literature), optimizing domain semantics, improving data quality by resolving inconsistencies and the completion of missing links.

2.2 Integration Strategies

In pursuit of knowledge, biologists constantly struggle with the massive amounts and scattered nature of the data they use. Integrating biological data therefore has always been one of the primary goals of bioinformatics. Several possible integration strategies can be used, as described by L. Stein [3]. *Link integration* resembles the fuzziness of the biological data it glues together. In Figure 1 a link integration scheme for microbial resource information is visualized. Biologists surfing the web reach related instances of biological data by clicking hyperlinks. Although cooperation between data sources is preferable, it is not deemed a strict necessity. There are however major disadvantages to this approach: *i*) There is no reduction of redundancy. Multiple copies of the same data can persist in the integration network and many mutual links need to be defined. *ii*) Missing links and data inconsistencies are hard to solve due to the complexity of this network.

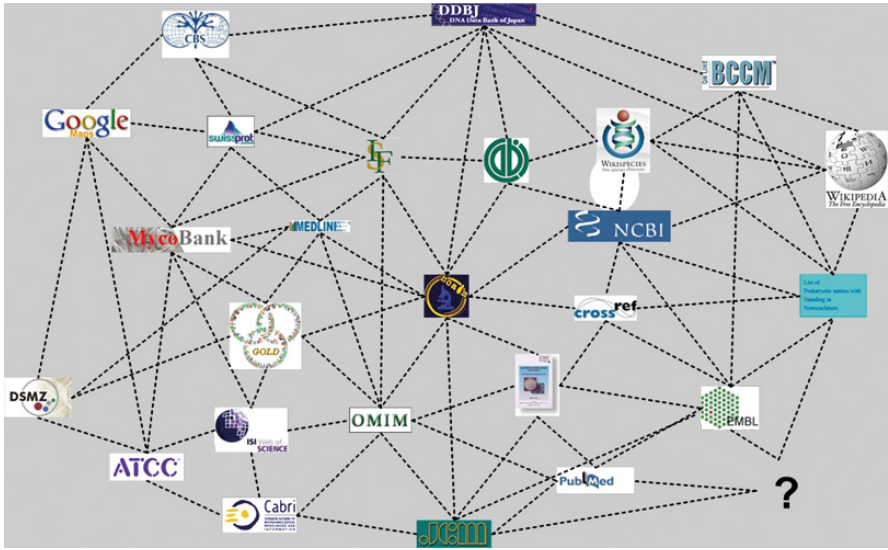


Fig. 1. Link integration scheme for microbiological resource data. Information providers (ATCC: the American Type Culture Collection, BCCM: the Belgian Coordinated Collection of Micro-organisms, CABRI: Common Access to Biological Resources and Information, CBS: Centraalbureau voor Schimmelculturen, CrossRef: <http://www.crossref.org>, DDBJ: DNA Databank of Japan, DSMZ: German Collection of Micro-organisms and Cell Cultures, EMBL: European Molecular Biology Laboratory, GOLD: Genomes OnLine, Google, ISI Web of Science, JCM: Japan Collection of Micro-organisms, Medline: <http://medline.cos.com/>, Mycobank: <http://www.mycobank.org>, NCAIM: National Collection of Agricultural and Industrial Micro-organisms, NCBI: the National Center for Biotechnology Information, OMIM: Online Mendelian Inheritance in Man, PubMed: <http://www.ncbi.nlm.nih.gov/books/bv.fcgi>, SwisSProt: www.expasy.org/sprot/, Wikipedia: <http://en.wikipedia.org>, Wikispecies: <http://species.wikimedia.org>) are linked in an unstructured and unregistered way. The question mark symbolizes the occurrence of *link rot*, where information locations on the web go bankrupt, creating voids in the linkage network.

Once an incorrect link is put in place, the network has a tendency to propagate the error without the availability of a centralised method ensuring data integrity. *iii*) When data providers change their internal system of accessing and naming records, external links may no longer be valid thus making link level integration very update dependent and vulnerable to ‘broken links’.

Another approach is *data warehousing* where data from multiple sources is fetched, transformed into a common data model and stored into the warehouse. Once the warehouse has been implemented, the original databases disappear for warehouse viewers. The main difficulties with this integration option comes with data synchronisation, standardisation and data fusion. When numerous databases have to be integrated, the same number of updates has to be performed on a regular basis and the warehouse designer has to be aware of any schema changes made in the source databases.

A hybrid approach combining the best of both link integration and data warehousing, as proposed by L. Stein [3], is the *knuckles-and-nodes* integration method. The setup of this strategy is relatively straightforward: a number of independent data providers are seen as highly specific and data-rich *nodes*. These are connected in a 1-to- n way to a *knuckle*, acting as an integration service by keeping track of updates and changes of its member nodes. It offers a common view on separate data objects while hiding the underlying specific technical details of the single nodes from its users. Emphasis lies on having the knuckles as lightweight as possible by minimizing data duplication and restrict local data content to the information needed for integration, quality control and backward linkage between the knuckle and its nodes. To avoid duplication of cross-references between objects, mutual links are maintained between knuckles, rather than between individual information nodes. This way, the network dynamics can be maintained following a true divide and conquer strategy. The next paragraph explains in more detail how the *knuckles-and-nodes* network of the StrainInfo.net biportal is implemented.

2.3 The StrainInfo.net Biportal Architecture

The current structure of the StrainInfo.net biportal is formed by four conceptual knuckles as clusters of independent data sources (nodes) containing microbial resource information. These knuckles represent independent domain objects that allow other knuckles and nodes to hook on. A visualization is given in Figure 2. The newly compiled specimen knuckle forms the cornerstone of the biportal and integrates the online catalogues of the BRCs, offering an interface to the information these centers provide on their strains. It uses globally unique identifiers to distinguish between its basic object instances, microbial cultures, while hiding the complexity of mutual relationships within the underlying BRC information network. This *culture identifier* is made publicly visible as it allows unified integration with third party databases. An important task of this knuckle is to solve ambiguities and inconsistencies in the synonymy of strain numbers [1]. It should be noted that the compilation of the sequence, literature and taxonomy knuckle, or any other knuckle representing a generalized information object in the microbial information network essentially falls outside the scope of the StrainInfo.net biportal. However, the StrainInfo.net biportal plays an important role in setting up cross-references between the specimen knuckle and the other knuckles established by third parties. The four knuckles and their mutual interactions are currently captured by the web interface of the StrainInfo.net biportal in five ‘views’. In the near future, a web service interface of the StrainInfo.net biportal will provide more detailed and flexible access to the knuckles and their mutual interactions.

Through the *synonyms* view of the portal’s web interface, a user is able to find all cultures for a given strain, including backward links to the corresponding online catalogue records of the BRCs that hold these cultures. This view also provides species information by making use of the specimen-taxonomy knuckle interface. The *map* view reveals the geographical location of all BRCs that have a

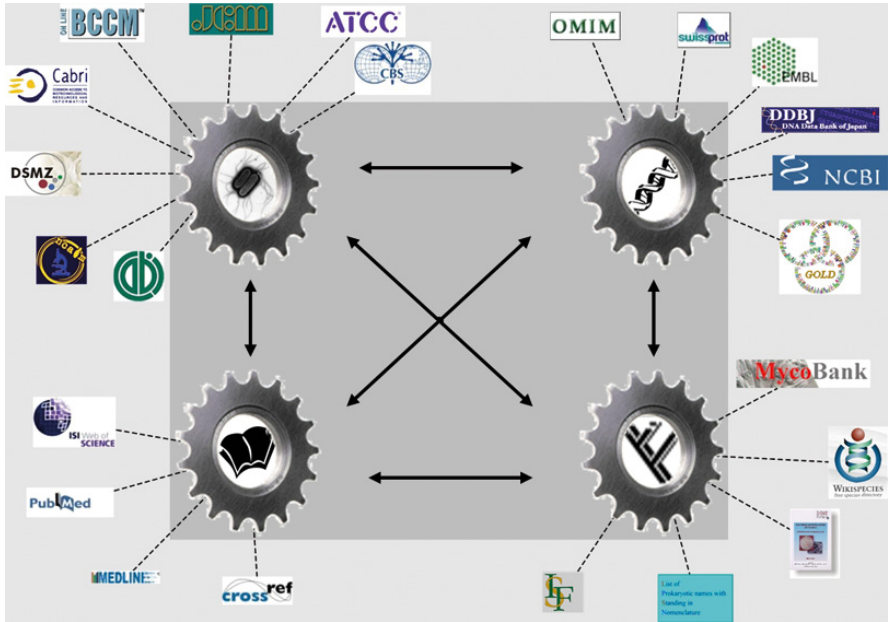


Fig. 2. Schematic design of the StrainInfo.net bioportal architecture. The specimen knuckle (upper left) groups all of the microbial resource information sources and forms the backbone of the bioportal. The StrainInfo.net bioportal maintains the interfaces between three other knuckles in the data space. The sequence knuckle (upper right) is a conceptualisation of sequence information. A literature knuckle (bottom left) forms the entry point to publication data while a taxonomy knuckle (bottom right) represents the taxonomic information available.

given strain in their holdings. The web interface to the specimen knuckle also incorporates code to graphically generate the strain history and presents it through the **history** view. This facilitates quality control on the involved cultures and can help dealing with legal issues on intellectual property rights. The interaction between the specimen knuckle and the sequence knuckle is implemented in the **sequence** view that provides links to the records of the sequences derived from a given strain. Similarly, the **literature** view shows all available literature information on a given strain.

The web interface of the StrainInfo.net bioportal thus forms a so-called one-stop-shop for microbial culture information. Researchers no longer need to browse numerous online BRC catalogues to retrieve complete information on a specific strain. But the real strength of the StrainInfo.net bioportal lies in the extensible interface it establishes between its specimen knuckle and other information knuckles. For data providers, the StrainInfo.net bioportal establishes a unique information network checking data integrity as will be discussed below. At the user side, researchers can find synonym cultures and the related sequence, taxonomic and literature information is just a click away without wor-

rying about the babel-like confusion that stems from the general practice of assigning multiple identifiers to the same biological material. Furthermore, the database underlying the StrainInfo.net bioportal provides maintenance and regular updates of this mutual information, while the web interface hides all of the underlying machinery.

3 Exploiting the Enriched Semantics of Integrated Microbial Information Networks

3.1 Integration at Work

Given the quantity of microbial data objects and their complex mutual links, putting all the pieces of the microbial information puzzle in place by hand is fairly impossible for human researchers. The StrainInfo.net bioportal aims at providing the research community with the most complete and correct integrated view on microbial information. The enriched semantics that follow from the integration process of the bioportal will be illustrated in the following examples.

A first example of the integration possibilities the StrainInfo.net bioportal offers is on the strain labelling of the specimen knuckle. As previously stated, BRCs use different identifiers for the same biological material. Knowledge about all these different identifiers is essential to perform searches in the different databases containing downstream information on the biological material. Manually fetching all the available synonyms of a strain would be a tremendous task, requiring *screen scraping* the cross-reference network originating from the identifiers of other collections or the `other collection numbers` fields that some, but not all, BRCs provide in their records.

Figure 3 graphically represents the catalogue cross-reference table for the *Pseudomonas stutzeri* type strain as gathered by the web spider running in the background of the StrainInfo.net bioportal. In this representation, the rows represent the different data source records that provide synonymy information on the identifiers assigned to the given strain. A black box in the table indicates that the corresponding data source mentions the corresponding strain identifier (self references are shown as gray boxes). For example the AS 1.1803 strain label used by the Institute of Microbiology at the Chinese Academy for Sciences is only referenced by the Japanese Collection of Microorganisms (JCM 5965) while other strain labels like LMG 2333 or ATCC 17588 are referenced by a larger number of data sources, thus having an increased visibility in the information network. This disequilibrium makes some labels easy to retrieve by human queries while it lets others tend to ‘get lost’ in the haystack of data. The StrainInfo.net bioportal overcomes this dependency on representation and incorporates all the known identifiers assigned to a strain when it receives one of its labels as a query input.

Besides integrating BRC data into the specimen knuckle, the StrainInfo.net bioportal also forms an entry point that provides access to the knuckle-to-knuckle network, that for example links bacterial sequence data to taxonomic information. An illustration of this can be found when investigating the AJ011504 Genbank record. The `source organism` field annotates this strain as *Pseudomonas*

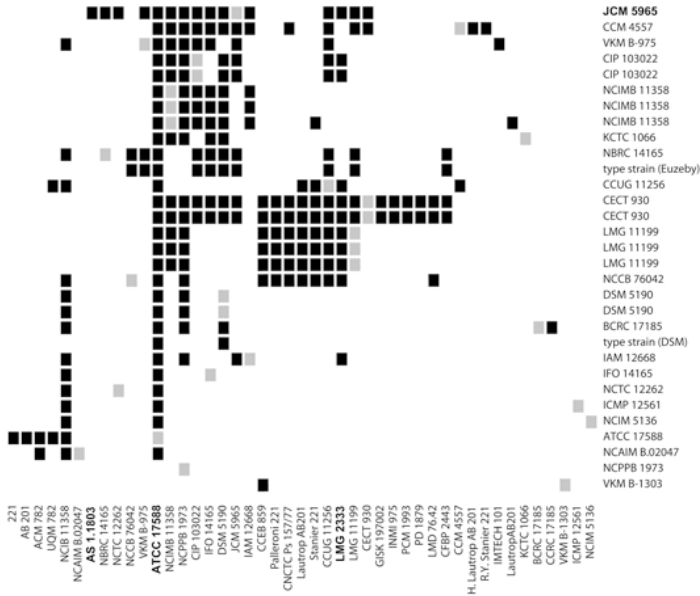


Fig. 3. Catalogue cross-reference table (CCRT) for synonym cultures as detected by the search engine of the StrainInfo.net bioportal. The rows represent the different data source records that provide synonymy information on the identifiers assigned to the given strain. A black box in the table indicates that the corresponding data source mentions the corresponding strain identifier (self references are shown as gray boxes). There’s a clear duality between strain labels showing a low degree of cross-referencing (e.g. AS 1.1803) and strain labels that are highly cross-referenced (e.g. LMG 2333, ATCC 17588).

sp., being an undefined species of the genus *Pseudomonas* and the strain field refers to the cultured labelled as BKME-9. However, this is a synonym of other identifiers as LMG 20220 and ATCC 700689, both taxonomically classified more accurately as *Pseudomonas abietaniphila* by their corresponding BRCs. Remark that the more accurate species identification has been applied in the AJ171416 record of the same sequence database, representing another sequences of the same biological material referenced using an alternative label (CIP 106708). To summarize the above inferences, Figure 4 shows how the *knuckles-and-nodes* architecture of the StrainInfo.net bioportal can be used to clarify some of the links in the data network. This way, the bioportal can indirectly provide the AJ011504 sequence with more accurate taxonomic information.

3.2 Distributed Intrusion Detection and Correction

When integrating multiple biological data sources, bioinformaticians must not only take into account the technological and conceptual issues. Besides getting the correct database drivers, writing the appropriate algorithms and designing the most suitable object schemas, the human interference with and the domain

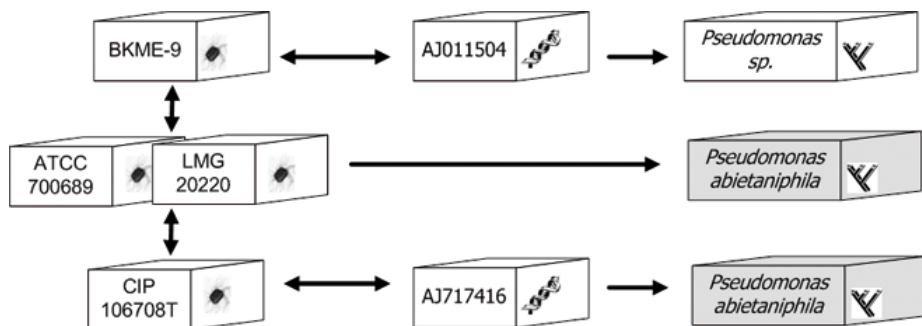


Fig. 4. Using the rich information network established by the StrainInfo.net bioportal to find the most accurate taxonomic annotation. The **source organism** field of the AJ011504 Genbank record annotates this strain as *Pseudomonas sp.* and the **strain** feature refers to the BKME-9 culture. However, this is a synonym of other identifiers as LMG 20220 and ATCC 700689, both taxonomically classified more accurately as *Pseudomonas abietaniphila* by their corresponding BRCs. Moreover, the **source organism** field of another sequence (AJ717416) from the same biological material, deduced from the synonym CIP 106708 label, also indicates that the sequence comes from a specimen that is taxonomically identified as *Pseudomonas abietaniphila*.

semantics behind the data must not be neglected. References between pieces of the biological knowledge are partially set by human researchers who unavoidably make unintended mistakes. As a result of this, errors might occur anywhere down the data stream. An efficient integration strategy must incorporate methods to find and – if possible – correct these inconsistencies.

Figure 5 shows an example of erroneous data originating from a somewhat hidden human error. The 16S rRNA sequence record with Genbank accession number X16895 references the strain number ATCC 12964. The associated BRC record declares this culture taxonomically as *Streptococcus pyogenes*. The Genbank file as well as the EMBL record, however, indicate that the species corresponding to this sequence is *Vibrio anguillarum*. From a taxonomic viewpoint *S. pyogenes* and *V. anguillarum* are clearly incompatible taxa.

The StrainInfo.net bioportal can solve this inconsistency, using its rich network of *knuckle-to-knuckle* interfaces. This is illustrated in Figure 6. By fetching the literature references of the sequence record through the sequence-literature knuckle interface, a publication can be found in which ATCC 19264 is mentioned instead of ATCC 12964. While scanning the taxonomy-specimen knuckle relationships, the same ATCC 19264 culture is identified as a type strain of *V. anguillarum*. Obviously, a simple orthographic error swapping two numbers in the strain identifier is responsible for the erroneous linking. This probably has happened during deposition of the X16895 sequence record in the public sequence database.

Detection of such inconsistencies can easily be automated by the bioportal. Semantic rules may be put forward and applied to the *knuckle-to-knuckle* network, providing a system by which the data space can be constantly monitored

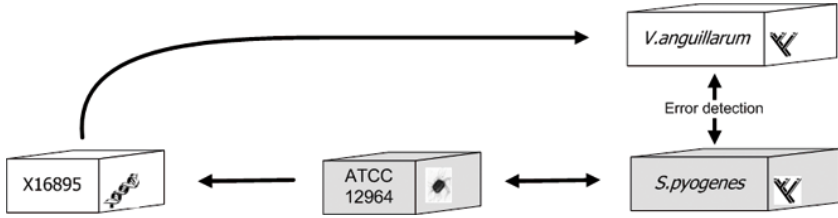


Fig. 5. Using the StrainInfo.net biportal network to detect data inconsistencies. The Genbank 16S rRNA sequence X16895 is referenced by the ATCC 12964 culture record. This BRC record declares the culture taxonomically as *Streptococcus pyogenes*. The sequence information provided by the public sequence databases, however, indicate that the species corresponding to this sequence is *Vibrio anguillarum*. From a taxonomic viewpoint *S. pyogenes* and *V. anguillarum* are completely incompatible taxa.

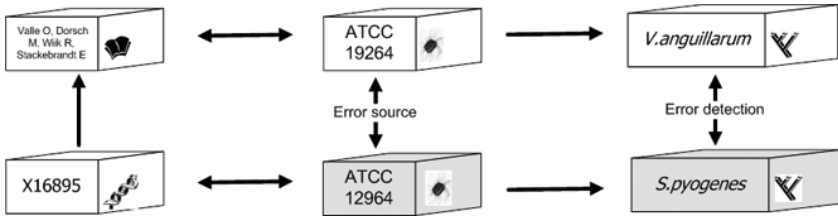


Fig. 6. The StrainInfo.net biportal can be used for more than the detection of errors. By making use of alternative connections in the rich information network, one could find the possible source of the error. In the ATCC 12964 example, taking a detour over the literature-specimen knuckle interface leads to the ATCC 19264 record, also assigned as *V. anguillarum*. Obviously, a typographic mistake between these two culture identifiers is the most plausible error source and probably happened on deposition of the X16895 sequence.

against unintended errors. Such rules could demand that the value of any attribute of an object is the same regardless of the fact that it was reached by path *x* or by path *y*. Although manual validation will be unavoidable, one may expect that the process of quality control might be largely automated in the near future.

4 Future Work

The current version of the StrainInfo.net biportal implements a classical web interface. This is not the endpoint of the project. In order to fully support the bioinformatics community [4], plans exist to implement a secondary interface layer of web services supporting more flexible computer agent interaction. This strategy of separate web services offering their functionality through XML messages and Simple Object Access Protocol (SOAP) is highly praised due to its component-based structure and independent nature [5]. As a consequence, even

when the internal procedures of a service change, as long as the interface remains the same, web service users do not need to modify their applications. The bioMOBY [6] and Taverna [7] projects are well-known examples of such web services systems.

The StrainInfo.net biportal can provide services towards the BRCs it integrates data from, whereas third party information providers can also benefit from the StrainInfo.net biportal. For instance, a major specific goal of the biportal is to establish a linkage system between strains and corresponding genetic sequence information [8]. Currently, mutual linkage between culture collection records and sequence information suffers from a number of problems. For instance, when depositing into the EMBL Data Library [9,10], strain information should be stored in the `isolate` or `strain` feature. However, depositors are not prohibited to omit or store this information in other fields. An example can be found in the CS302340 EMBL entry where the strain name is given in the `organism name` field instead of the `strain` feature.

Furthermore, since there are no strict naming rules, strain identifiers not following the acronym-space-alphanumeric identifier convention are also hard to filter out of the flat files of the sequence database by parsing for strain information. Besides these shortcomings, allowing the same biological material to have multiple labels raises the issue of 'ID disparity' where one strain has multiple IDs in different databases, duplicate IDs in the same database and erroneous IDs everywhere. Moreover, all of these errors tend to propagate very easily to several related databases. As biological resources are essential tools for biological research, a more efficient linking of strain information to sequence databases could improve access to micro-organisms of certified quality and be beneficial to the sequence databases by updating their records based on the sequence-specimen knuckle interface of the StrainInfo.net biportal. Therefore, the biportal will attempt to set up a solid cross-reference scheme in cooperation with the major sequence databases.

Examples of other data providers that may possibly hook themselves upon the StrainInfo.net biportal are the recent Integrated Microbial Genomes system [11] and the KEGG pathway database [12].

5 Conclusions

The StrainInfo.net biportal envisions to overcome some of the problems related to the integration of basic information on biological resources. Besides dealing with the heterogeneous nature of data provided by hundreds of BRCs worldwide, the StrainInfo.net biportal wants to accept the challenge of integrating the dynamically growing amount of downstream information on these organisms in a single biportal. Though the specimen knuckle forms the solid rock base of the project, the StrainInfo.net biportal wants to play a pronounced role in a mutual linkage scheme between strain information and third party knowledge repositories. At first, the services of the biportal will be available through a web interface, but plans are at hand for the development of community support by web services.

Acknowledgements

Special thanks go to Prof. Juncai Ma and Mr. Xianhua Zhou, from the Information Network Center, Institute of Microbiology, Chinese Academy of Sciences in Beijing. Their efforts on the development of an initial prototype version of the StrainInfo.net bioportal has greatly inspired the compilation of the current second prototype version. Prof. Dr. Paul De Vos and Prof. Dr. Bernard De Baets are indebted to the Belgian Ministry of Science Policy for grant C3/00/12.

References

1. Dawyndt, P., Vancanneyt, M., De Meyer, H., Swings, J.: Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Transactions on Knowledge and Data Engineering* **17** (2005) 1111–1126
2. CABRI: Common Access to Biological Resources and Information. (<http://www.cabri.org/>)
3. Stein, L.: Integrating biological databases. *Nature Reviews Genetics* **4** (2003) 337–345
4. Stein, L.: Creating a bioinformatics nation. *Nature* **417** (2002) 119–120
5. Gao, H., Huffman Hayes, J., Cai, H.: Integrating biological research through web services. *Computer* **38** (2005) 26–31
6. Wilkinson, M.D., Links, M.: BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics* **3** (2002) 331–341
7. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M., Wipat, A., Li, P.: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20** (2004) 3045–3054
8. Romano, P., Dawyndt, P., Piersigilli, F., Swings, J.: Improving interoperability between microbial information and sequence databases. *BMC Bioinformatics (Suppl 4)* (2005) S23
9. Emmert, D., Stoehr, P., Stoesser, G., Cameron, G.: The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Research* **26** (1994) 3445–3449
10. Stoesser, G., Sterk, P. and Tuli, M., Stoehr, P., Cameron, G.N.: The EMBL Nucleotide Sequence Database. *Nucleic Acids Research* **25** (1997) 7–13
11. Markowitz, V., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N., Kyrpides, N.: The integrated microbial genomes (IMG) system. *Nucleic Acids Research* **34** (2006) D344–D348
12. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. **30** (2002) 42–46

MIOS-CIAO 2006 PC Co-chairs' Message

Modern enterprises face a strong economical pressure to increase competitiveness, to operate in a global market, and to engage in alliances of several kinds. Agility has thus become the new guiding principle for enterprises. This requires flexible organizational structures and business processes, as well as flexible supporting information systems and a flexible information and communication technology (ICT) infrastructure. In addition, an enterprise needs to be able to easily expand or shrink, be it through collaborations with other enterprises, through mergers or acquisitions, or through insourcing or outsourcing of services.

In order to meet these economical requirements, enterprises rely increasingly on the benefits of modern ICT. However, the appropriate knowledge to deploy this technology as needed, and in an effective and efficient way, is largely lacking, particularly knowledge regarding the cooperation in and between enterprises and knowledge regarding the interoperability of their information systems.

Up to last year, several successful MIOS workshops (Modeling Inter-Organizational Systems) were held. This year the main focus of the workshop was set on CIAO (Cooperation and Interoperability, Architectures and Ontologies). We hope you will enjoy reading these papers and you will find them valuable for your research and knowledge.

August 2006

Antonia Albani, University of Augsburg, Germany
Jan L.G. Dietz, Delft University of Technology, Netherlands

The CrocodileAgent: Research for Efficient Agent-Based Cross-Enterprise Processes

Vedran Podobnik, Ana Petric, and Gordan Jezic

University of Zagreb, Faculty of Electrical Engineering and Computing,
Department of Telecommunications, Unska 3, HR-10000 Zagreb, Croatia
{vedran.podobnik, ana.petric, gordan.jezic}@fer.hr

Abstract. The Trading Agent Competition (TAC) is an international forum which promotes high-quality research regarding the trading agent problem. One of the TAC competitive scenarios is Supply Chain Management (SCM) where six agents compete by buying components, assembling PCs from these components and selling the manufactured PCs to customers. The idea of TAC SCM is the development of agent-based architectures that implement wide-applicable business strategies which efficiently manage cross-enterprise processes. In this paper, we analyze the TAC SCM environment and describe the main features of the CrocodileAgent, our entry in the TAC SCM Competition.

1 Introduction

The initial architecture of the Web was geared towards delivering information visually to humans. We are currently witnessing a transformation in the architecture of the Internet which is becoming reposed on goal directed applications that intelligibly and adaptively coordinate information exchanges and actions¹. Consequently, the Internet is transforming into an enabler of the digital economy. The digital economy, by proliferation of the use of the Internet, provides a new level and form of connectivity among multiple heterogeneous ideas and actors, giving rise to a vast new range of business combinations [1]. Additionally, by utilizing the technology of intelligent software agents, the digital economy automates business transactions.

Everyday business transactions generate an enormous quantity of data which often contains a lot of information about business activities that often goes unnoticed. A systematic approach to the available data (proper data preparation and storage) can assure well-timed and well-placed high-quality information. An integrated enterprise information system is an imperative infrastructure precondition for high-quality information. The role of such a system is to unite all company parts and functions into a single entity whose duty is to completely satisfy all the company's information needs. This system consists of an adequate database (a data warehouse) and business applications. Business applications are nowadays usually divided into three parts - ERP (*Enterprise Resource Planning*), SCM (*Supply Chain Management*) and CRM (*Customer Resource Management*). EAI (*Enterprise Application Integration*) is a

¹ Source: IBM.

common term used for these applications integrated together. ERP is an operating information system which integrates and automates company business activities. Consequently, ERP is a supporting system for SCM and CRM such that it enables intra-enterprise collaboration. The SCM is an automated system consisting of processes and procedures that are in charge of interaction with company *suppliers*. The CRM is, on the other hand, an automated system consisting of processes and procedures that are in charge of interaction with company *customers*. Therefore, the SCM and the CRM are systems that enable inter-enterprise collaboration.

In today's economy, supply chains are still based on static long-term relationships between trading partners. These relationships are the main obstacle in realising dynamic supply chains where the market is the driving force. Dynamic supply chain management improves the competitiveness of companies since it has a direct impact on their capability of adjusting to the changing market demands quickly and efficiently [2]. Since the annual worldwide supply chain transactions are counted in trillions of dollars, even the slightest possibility of improvement cannot be neglected.²

The purpose of the TAC SCM (*Trading Agent Competition Supply Chain Management*) game is to explore how to maximize the profit in the stochastic environment of volatile market conditions. Thus, it is important to develop an agent capable of reacting quickly to changes taking place during the game. Furthermore, it is critical to implement predictive mechanisms which enable an agent's proactive behaviour. Although the name of the game is TAC SCM, it does not only deal with problems of systems that are nowadays known as the SCM systems. Namely, it requires participants to investigate autonomous solutions for managing complete EAI systems. The idea is to build a robust, highly-adaptable and easily-configurable mechanism for efficiently dealing with all EAI facets, from material procurement and inventory management to goods production and shipment [3]. Additionally, TAC SCM tournaments provide an opportunity to analyze effects common in real-world business transacting, such as the bullwhip effect, and its relationship with company profits [4]. Furthermore, the tournament can help in developing methods for identifying the current economic regime and forecasting market changes [5].

In this paper, we analyze the TAC SCM environment and describe the main features of the CrocodileAgent, an intelligent agent we developed to participate in the TAC SCM Competition. The paper is organized as follows. Section 2 describes the basic rules of the TAC SCM game. Section 3 describes the CrocodileAgent's architecture, functionalities and its performance in the TAC SCM 2006 Competition. Section 4 proposes directions for future work and concludes the paper.

2 The TAC SCM Game

The connection between AI (*Artificial Intelligence*) and economics has received a lot of attention recently [6]. The TAC SCM game is also based on that connection. The Trading Agent Competition (TAC, www.sics.se/tac) is an international forum that

² Here we primarily observe markets from the price point of view, what is valid for TAC SCM environment. If we want to analyze the real world markets, aspects such as product quality and trust should also be taken into account.

promotes high-quality research on the trading agent problem. TAC has two competitive scenarios. The older one is called the TAC Classic scenario where eight agents compete by assembling travel packages for customers with different preferences for their trip. The other one is the TAC SCM game [7] that started in 2003. After four annual TAC SCM tournaments, intra-tournament market efficiency significantly increased [4].

In the TAC SCM game [8] scenario, each of six agents included in the game has its own PC (*Personal Computer*) manufacturing company. During the 220 TAC SCM days (one virtual day lasts 15 seconds), agents compete in two different markets (Figure 1). In the B2B e-market agents compete in buying PC components necessary to produce PCs. Participants in that market are the agents and eight suppliers. Each supplier produces four types of components (CPUs, motherboards, memories and hard drives) with different performance measures. Each agent has its own PC assembling factory with limited capacity. In each factory, 16 different types of PCs can be manufactured. The PCs are divided into three market segments: low range, mid range and high range (different ranges are characterized with different customer demand levels). In the B2C e-market, agents try to sell all the PCs they produced to customers and, at the same time, earn as much money as possible. The winner is the TAC SCM agent with the most money in its bank account at the end of the game.

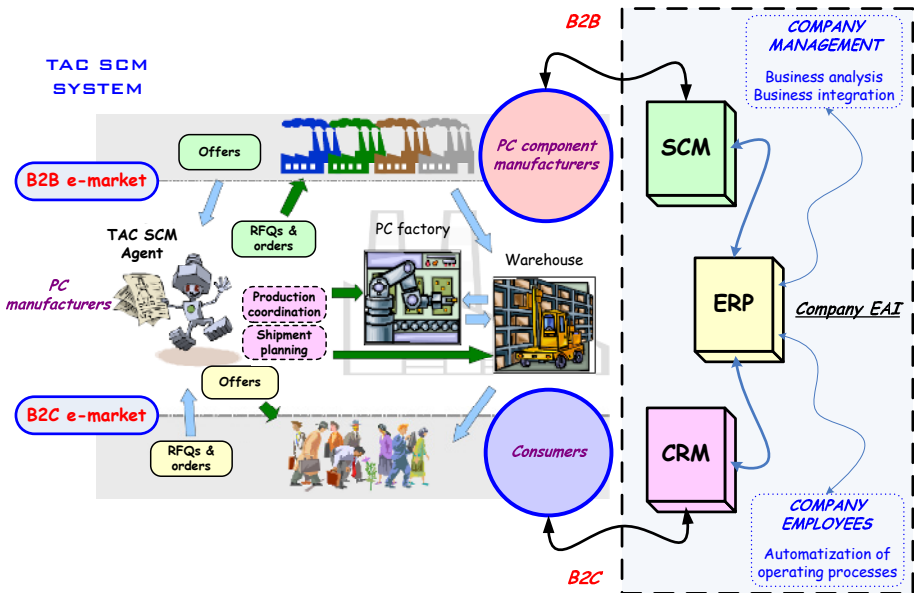


Fig. 1. The relationship between the TAC SCM system and a real company EAI system

The architecture of the TAC SCM system is shown in Figure 2. In order to participate in the game, an agent has to connect to the game server. The TAC SCM game server has several functionalities. Namely, it simulates suppliers (PC component manufacturers), customers (PC buyers) and the bank. The game server

also controls agents' factories and warehouses. Each TAC SCM agent has a bank account and receives a daily report regarding its current bank balance. At the beginning of the game, the agent has no money and must hence loan money from the bank. For every day that the agent is in debt, the bank charges the agent interest while for every day that its bank account is positive, the bank pays interest to the agent.

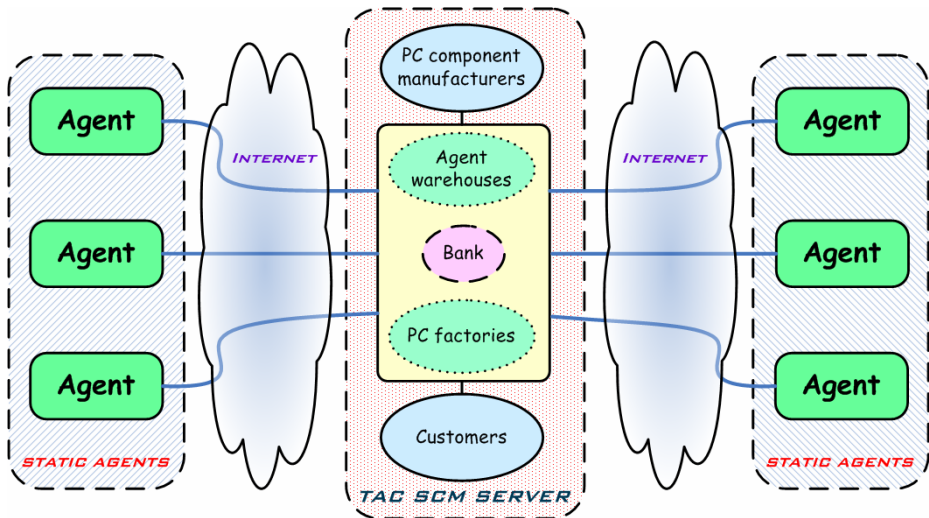


Fig. 2. The architecture of the TAC SCM system

3 The CrocodileAgent

3.1 Intelligent Software Agents as Enablers of the Digital Economy

An intelligent software agent is a program which acts on behalf of its owner while conducting complex information and communication actions over the Web. Figure 3 shows the relations between the main features of intelligent software agents [9, 10].

The features of the technology of intelligent software agents make them perfectly applicable in modern enterprise systems and electronic markets (e-markets). In the past, both the markets and choices available were much smaller than today, so the volatility of supply and demand functions was much more inert. Under such market conditions, companies did not need to make important decisions daily but they rather based business transactions on long-term partnerships. The accelerated economic globalization trend in the past decade is leading us closer to the existence of just one market - the global one. Consequently, the functions of supply and demand are becoming more and more dynamic and the possibilities of choice have risen to amazing levels. This is a reason why companies today have great difficulties in enhancing the efficiency of their current business processes. Companies are instantly forced to make lots of important decisions while continuously trying to maximize their profits. Keeping in mind the great volatility that characterizes the complex set of

market conditions and the vast quantity of available information, a possible solution for improving business efficiency is the automation of business processes and excluding humans from making decisions (where this is possible). Humans simply do not possess the cognitive ability to process such an enormous quantity of information (and to make adequate decisions) in the few moments during which the relevant information does not change. A very logical solution to this problem lies in the technology of intelligent software agents – i.e. computer programs with the ability to completely autonomously manage a set of tasks.

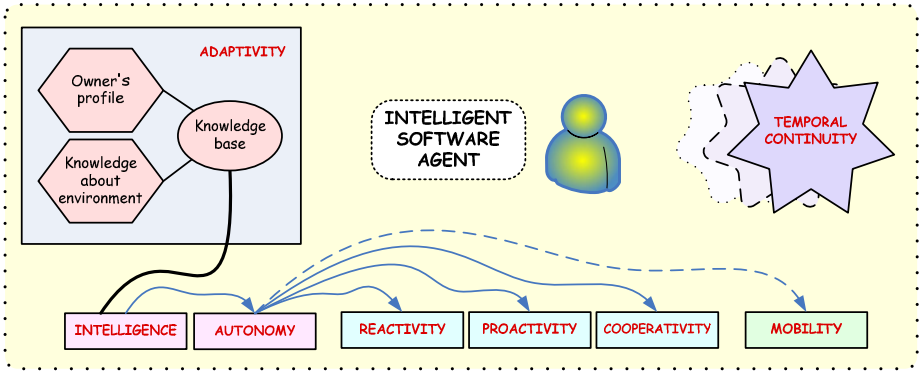


Fig. 3. A model of intelligent software agent

3.2 The CrocodileAgent’s Architecture

The CrocodileAgent [11, 12] is an intelligent agent developed at the Department of Telecommunications, Faculty of Electrical Engineering and Computing in Zagreb, Croatia. The CrocodileAgent's architecture (presented in Figure 4) is based on the IKB model [13], a three layered agent-based framework for designing strategies in electronic trading markets.

The first layer is the Information Layer (IL) which contains data gathered from the ongoing game. This data can be divided into two parts: data gathered from the market which is available to all agents and the private data about the agent's actions in the game. The IL also contains data about past games. Due to a limited capacity, the IL can not contain all the information available about the game. The information that is stored into the IL is determined by the Information Filter (IF).

The second layer is the Knowledge Layer (KL) which represents the knowledge acquired from the data stored in the IL. The KL determines and modifies the settings of the IF. Knowledge contained in the KL can be divided into knowledge of the agent's state and the market's state.

The third layer is the Behavioural Layer (BL) which is a decision-making component that determines the agent's strategic behaviour. The BL uses knowledge from the KL to make important decisions regarding component purchases, the production schedule, PC sales and shipment.

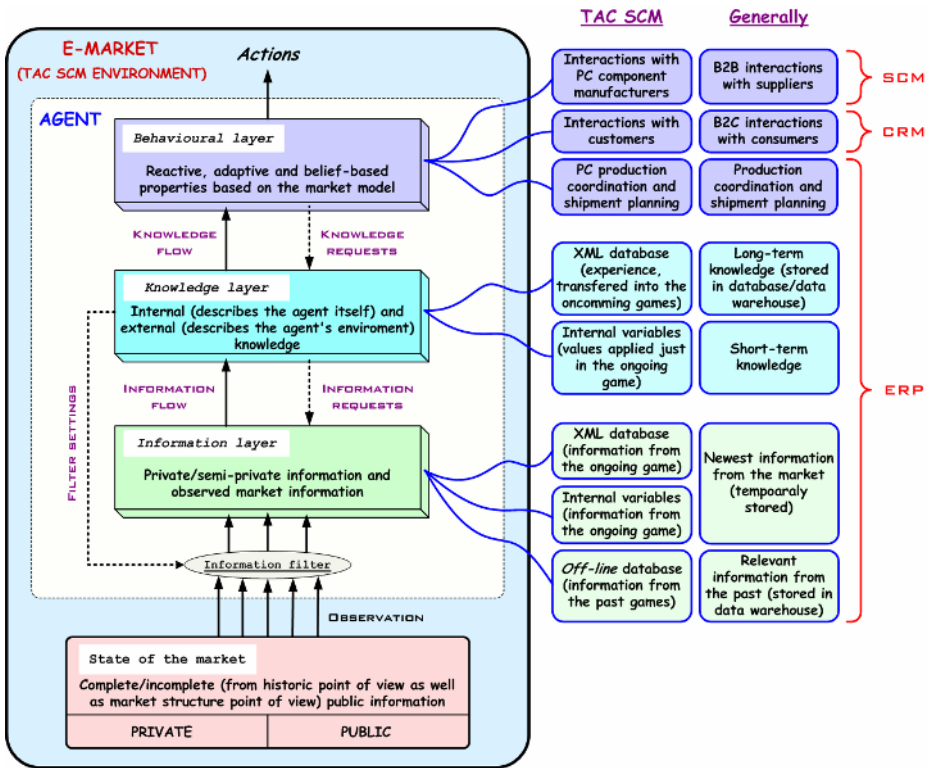


Fig. 4. The CrocodileAgent’s architecture

3.3 The CrocodileAgent’s Mechanisms for Managing Cross-Enterprise Processes

Enterprise Resource Planning (ERP). Initially the CrocodileAgent produced PCs only after receiving customer orders, but later we added the possibility of producing PCs even if nobody ordered them. Since the TAC SCM game is stochastic in nature, customer demand varies during the game. If the agent does not produce PCs and the PC demand is low, a large part of the agent’s factory capacities remain unutilized. If the agent produces PC stocks during a period of low PC demand, its factory will be utilized and the agent will be prepared for a period of high PC demand. This tactic has its weaknesses since the TAC SCM agent cannot know for sure what the future demand will be. Hence, the agent might produce the stock of PCs that cannot be sold for a longer period of time. The CrocodileAgent lowers this risk by introducing quantity limits which represent the maximum number of PCs which can be available in stock, for every PC type separately. Quantity limits are adjusted depending on the predicted level of future demand.

Every day the list of active orders is sorted chronologically according to the delivery dates and then an algorithm for PC production and shipment to customers is executed. The algorithm runs as follows:

- The agent checks if there are enough PCs in the warehouse to fulfill the order
 - If so, they are reserved and added to the delivery schedule.
 - Otherwise, the agent checks if there are enough components to produce the requested PCs
 - ✗ If so, the components are reserved and the agent tries to add them to the production schedule. The production demand will be successfully fulfilled only if there is enough free factory capacity available for the next day.

After analyzing all the active orders, the CrocodileAgent checks the production schedule for the next day. If there is free capacity available, the agent uses it for creating the allowed level of PC stocks.

Supply Chain Management (SCM). The CrocodileAgent grounds its procurement strategy on short-term purchasing of smaller quantities of PC components. Using such a strategy prevents the agent from paying large amounts of money to stock PC components in their warehouse.

RFQs (*Requests for Quotes*) are sent to all the suppliers that produce the needed component. The agent's accepts only the cheaper offer, which causes a temporary degradation of the agent's reputation in the eyes of the supplier whose offer was not accepted. However, since the requested quantities are not high, the agent's reputation quickly recovers. The agent accepts partial offers in case the chosen supplier cannot deliver the requested quantity on time. This way, the agent gets a smaller amount of components than planned, but on time. In that case the quantities and reserve prices³ are modified for more aggressive purchasing of that component in the next few days.

Some of the parameters used in component purchase are:

- N_{\min} – the minimal quantity of components required to be in storage;
- N_{\max} – the maximal quantity of components allowed in storage;
- N_{ord} – the maximal amount of components that can be ordered in one day;
- N_{res} – the quantity of a component reserved for PC production for the next day;
- N_{inv} – the number of components currently stored in the warehouse.

At the beginning of each day, the agent calculates the component quantity ordered, but not delivered, up to that moment for each component separately. The CrocodileAgent's ordered quantities of components are multiplied with a distance factor. The distance factor is a value between 0 and 1; the factor shrinks from 1 to 0 as the delivery date grows. When the delivery date reaches 30 days (from the current day) the distance factor becomes 0. The parameter obtained by performing this calculation is referred to as the *evaluatedQuantity*. Similarly, we calculate the *evaluatedLongTermQuantity* which represents the quantity of all the ordered components that have a delivery date higher than 30 days.

For each component, the agent checks to determine if the following condition is fulfilled:

$$N_{\text{inv}} + \text{evaluatedQuantity} > N_{\text{max}} . \quad (1)$$

If condition (1) is fulfilled, the agent does not order this component, but counts the number of days in a row that the component is not ordered. If the agent does not order

³ Reserve price is the maximum price that the agent is willing to pay for the component.

components for five days in a row and if the *evaluatedLongTermQuantity* is under its upper limit in spite of condition (1), the agent makes long-term orders to ensure cheap components later in the game. If condition (1) is not fulfilled, the following condition is considered:

$$N_{inv} + N_{res} > N_{min}. \quad (2)$$

If condition (2) is fulfilled, the agent sends three short-term RFQs with the purpose of maintaining the present level of components in the warehouse. In case condition (2) is not fulfilled, the agent also sends three short-term RFQs but with the purpose of getting the number of components in the warehouse above N_{min} as soon as possible. Regardless of condition (2), the agent sends two long-term RFQs to ensure long-term occupancy of the warehouse. It is important to point out that these are only the main characteristics of the algorithm. Additionally, there are special mechanisms which calculate the reserve prices and exact quantities that need to be ordered. A description of two of these mechanisms follows:

- The *lowComponentAlarm* marks the very low quantity of a certain component in the warehouse. It allows short-term procurement of this component where the agent pays a higher price than usual. Since different components have different prices, the reserve price for them also differs, e.g. if the *lowComponentAlarm* is set, the maximal reserve price for processors (which are the most expensive PC component) is 115% of their nominal price and for other components it is 130% of their nominal price.
- The *demandPurchaseQuantityFactor* is modified according to customer demand. If the demand rises rapidly, the agent uses more components to produce more PCs, so the parameter is increased to ensure that the agent does not run out of components as a result of increased PC demand.

Customer Relationship Management (CRM). This aspect of the TAC SCM agent is responsible for sending offers to customers as replies to their RFQs. One agent, due to its limited factory capacity, is not able to produce PCs for all the RFQs issued in one day. Moreover, six TAC SCM agents compete for every RFQ and just the one with the lowest bid price will win the order. Therefore, the CrocodileAgent must carefully choose to which RFQs to reply and what prices to offer [14, 15].

Each day the CrocodileAgent first calculates the production cost of every PC type by summing the average purchase prices of each component incorporated in that PC. If some component type is not used in PC production for several days in a row, this usually means that it was purchased at a high price which is no longer concurrent on the market. As a result, the agent puts a discount on it. This way the agent prevents a further blockade of selling PCs which contain the expensive component. The discount grows as the period of component inactivity is longer.

Furthermore, the CrocodileAgent daily calculates its minimal profit per offer. Its profit margin depends on the overall level of customer demand (greater demand yields a larger profit margin), the factory capacity reserved for active orders (if most of the factory capacity in the next few days is reserved, the profit margin increases) and the due date listed in the RFQ (the sooner is the due date, the higher the profit).

After sorting the RFQs in descending order for every PC type separately according to reserve prices customers are willing to pay, the agent starts to send offers if two conditions are fulfilled:

- There are enough unreserved components (or already produced and stocked unreserved PCs) in the warehouse for producing the PCs requested in currently processed RFQ. The CrocodileAgent reserves components (or already produced PCs if they exist) for every offer it sends. Actually, it does not reserve the complete quantity, but that quantity multiplied by the targeted offer acceptance rate (explained later in Figure 5).
- The price obtained by summing agent's PC production cost and agent's profit margin is lower than the customer's reserve PC price.

This algorithm comes in two versions: *handleCustomerRFQsNormal* and *handleCustomerRFQsHighDemand*. The version that is active on a certain day is determined depending on the number of production cycles needed to produce all the active orders and the algorithm that was used the day before. The basic difference between these two algorithms is the method of determining the offer prices for the PCs. The most frequently used algorithm during the game is *handleCustomerRFQsNormal* that determines the offer price by applying the model described later in Figure 5. The algorithm *handleCustomerRFQsHighDemand* is a "greedy" algorithm since the offer prices for the PCs are always just slightly under the customer's reserve price. It is used when there is a very high customer demand for PCs since, in those cases, agents usually do not send offers for all the RFQs received. After analyzing the RFQs that did not get any offers, we noticed that some of them were very profitable. As a result, we decided to send offers for them but with high offer prices. This algorithm is also used at times when the agent's factory does not have much free capacity in the next few days. This happens when the agent receives a lot of orders. By sending offers with very high offer prices, the agent will not receive many orders. This prevents the situation in which the agent cannot deliver all the ordered PCs, and at the same time, a high profit is achieved with the few orders the agent wins.

Both versions of the selling algorithm implement a mechanism used to prevent late deliveries and keep the agent from paying penalties. Each day, the agent monitors its obligations to customers by calculating the number of factory cycles needed to fulfill its existing orders for each delivery date. After analyzing these numbers, the CrocodileAgent calculates the closest possible delivery date for the new orders and does not reply to RFQs with earlier due dates. This way the agent is prevented from sending offers that cannot be delivered by the requested delivery date.

Initially, the CrocodileAgent reserved components when it sent offers to customers and only considered making new offers if it had enough unreserved components to produce the requested PCs. Since the agent does not receive orders for all the offers it sends, some of the components stay unused for a long time. This problem was solved by introducing a prediction mechanism which calculates the price interval for assuring the desired offer acceptance rate. The TAC SCM agent knows the minimal and maximal price of every PC type from transactions on the previous day. The CrocodileAgent uses this information and by applying linear regression it determines the targeted price interval (described in Figure 5).

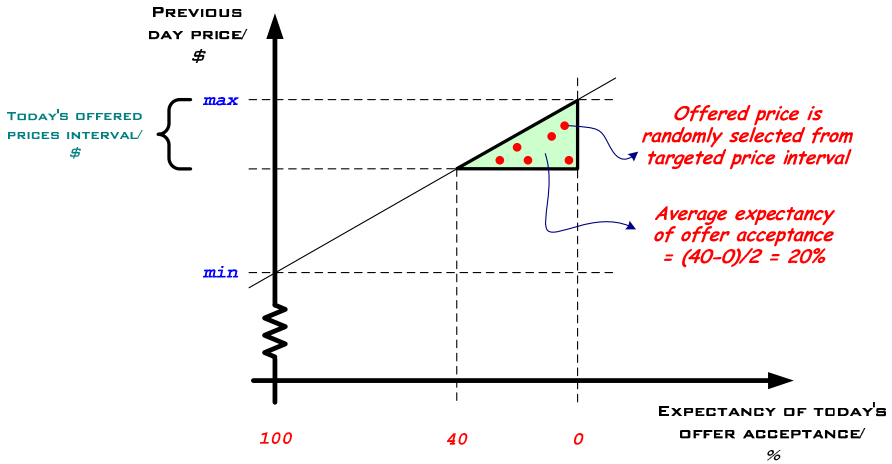


Fig. 5. A model for determination of offered prices

3.4 TAC SCM 2006 Competition

The CrocodileAgent participated in TAC SCM Competitions in years 2004, 2005 and 2006. It reached the semi-finals once and was twice eliminated in the quarter-finals.

The 2006 TAC SCM Competition (there were 23 teams competing, representing universities from all around the world) was divided into three parts: the qualifying round held from April 3rd – 11th, the seeding rounds held from April 24th – May 2nd and the final round held from May 9th – 11th. The CrocodileAgent's final score after the qualifying round was 8.275M which was enough to place the agent in 11th place. Our TAC SCM agent ended the seeding rounds at 11th place with an average score of 6.332M. The CrocodileAgent took 4th place in the quarter-finals with an average score of 8.470M and ended its participation in 2006 TAC SCM Competition.

4 Conclusions and Future Work

In this paper we presented the TAC SCM Competition and described the main features of the CrocodileAgent, our entry in that competition. The idea of TAC SCM is the development of agent-based architectures that implement wide-applicable business strategies which efficiently manage cross-enterprise processes. Therefore, the TAC SCM Competition is used as a benchmark for various solutions to this problem. We analyzed the TAC SCM environment and explained why agent-based systems are very applicable for electronic trading and automation of business processes. In addition, we highlighted some characteristic problems in the automation of cross-enterprise processes and proposed possible solutions that were implemented in our agent.

For the TAC SCM 2007 Competition we plan to upgrade the predictive mechanisms by using machine learning techniques (i.e. decision trees) and increase the CrocodileAgent's reactivity with a fuzzy logic approach [16].

References

1. Carlson, B.: The Digital Economy – What is New and What is Not?. Structural Change and Economic Dynamics, Vol. 15, Elsevier (2004). 245-264
2. Benish, M., Sardinha, A., Andrews, J., Sadeh, N.: CMieux – Adaptive Strategies for Competitive Supply Chain Trading. In Proc. Of the 8th Int. Conference on Electronic Commerce (ICEC), Fredericton, Canada, 2006.
3. Kontogounis, I., Chatzidimitriou, K.C., Symeonidis, A.L., Mitkas, P.A.: A Robust Agent Design for Dynamic SCM Environments. In Proc. Of the 4th Hellenic Joint Conference on Artificial Intelligence (SETN), Heraklion, Greece, 2006. 127-136
4. Jordan, P.R., Kiekintveld, C., Miller, J., Wellman, M.P.: Market Efficiency, Sales Competition, and the Bullwhip Effect in the TAC SCM Tournaments. In Proc. of the AAMAS Joint Int. Workshop on the Trading Agent Design and Analysis and Agent Mediated Electronic Commerce (TADA/AMEC) Hakodate, Japan, 2006. 99-111
5. Ketter, W., Collins, J., Gini, M., Gupta, A., Shrater, P.: Identifying and Forecasting Economic Regimes in TAC SCM. In Proc. of the IJCAI Workshop on Trading Agent Design and Analysis (TADA), Edinburgh, UK, 2005. 53-60
6. Wurman, P.R., Wellman, M.P., Walsch, W.E.: Specifying Rules for Electronic Auctions. AI Magazine, Vol. 23 (3), American Association for Artificial Intelligence (2002). 15-24
7. Eriksson, J., Finne, N., Janson, S.: Evolution of a Supply Chain Management Game for the Trading Agent Competition. AI Communications, Vol. 19, IOS Press (2006). 1-12
8. Collins, J., Arunachalam, R., Sadeh, N., Eriksson, J., Finne, N., Janson, S.: The Supply Chain Management Game for the 2006 Trading Agent Competition. http://www.sics.se/tac/tac06scmspec_v16.pdf. Date accessed: July 7, 2006.
9. Bradshaw, J.M.: Software Agents. MIT Press, Cambridge, Massachusetts, USA (1997)
10. Chorafas, D.N.: Agent Technology Handbook. McGraw-Hill, New York, USA (1998)
11. Petric, A., Jurasovic, K.: KrokodilAgent: A Supply Chain Management Agent. In Proc. of the 8th Int. Conference on Telecommunications (ConTEL), Zagreb, Croatia, 2005. 297-302
12. Petric, A., Podobnik, V., Jezic, G.: The CrocodileAgent: Analysis and Comparison with Other TAC SCM 2005 Agents. In Proc. of the AAMAS Joint Int. Workshop on the Trading Agent Design and Analysis and Agent Mediated Electronic Commerce (TADA/AMEC) Hakodate, Japan, 2006. 202-205
13. Vytelingum, P., Dash, R.K., He, M., Jennings, N.R.: A Framework for Designing Strategies for Trading Agents. In Proc. of the IJCAI Workshop on Trading Agent Design and Analysis (TADA), Edinburgh, UK, 2005. 7-13
14. Pardoe, D., Stone, P.: Bidding for Customer Orders in TAC SCM: A Learning Approach. In Proc. of the AAMAS Int. Workshop on Trading Agent Design and Analysis (TADA), New York, USA, 2004.
15. Burke, D.A., Brown, K.N., Tarim, S.A., Hnich, B.: Learning Market Prices for a Real-time Supply Chain Management Trading Agent. In Proc. of the AAMAS Joint Int. Workshop on the Trading Agent Design and Analysis and Agent Mediated Electronic Commerce (TADA/AMEC) Hakodate, Japan, 2006. 29-42
16. He, M., Rogers, A., Luo, X., Jennings, N.R.: Designing a Successful Trading Agent for Supply Chain Management. In Proc. Of the 5th Int. Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS), Hakodate, Japan, 2006. 1159–1166

Case Study – Automating Direct Banking Customer Service Processes with Service Oriented Architecture

Andreas Eberhardt¹, Oliver Gausmann², and Antonia Albani³

¹ b.telligent GmbH & Co.KG
85748 Garching/Munich, Germany
andreas.eberhardt@btelligent.de

² University of Augsburg
Chair of Business Informatics and Systems Engineering
86159 Augsburg, Germany

oliver.gausmann@wiwi.uni-augsburg.de

³ Delft University of Technology
Information Systems Design
2628 CD Delft, The Netherlands
a.albani@tudelft.nl

Abstract. The direct banking business is characterized by integrated distribution channel politics and holistic sales approaches combined with multi-channel-management. Direct banks in Europe and especially in Germany are currently facing increasing market competition. The crucial factors for growth are product innovation, cost control and the flexibility to react individually to each customer in a rapidly changing business environment. In order to compete, direct banks are forced to undergo a drastic transformation of business processes as well as organizational and managerial structures. The application of new concepts in building information systems is therefore necessary in order to further support business needs and allow for the management and adaptation of systems that are dependent on the fast changing market requirements. This paper shows how the information technology (IT) landscape of one of the five leading direct banks in Germany could be optimized by means of a service-based orientation. The case outlined in this paper focuses on the customer service domain. The main goal is to concurrently reduce costs by automating business processes and to increase the quality of customer services. A reference model for these customer service processes is then introduced. Based on this model, this paper describes a business component-oriented system architecture according to identified business components, and their corresponding services.

1 Introduction

Foreign banks entering the German banking market and the development of new customer specific products are putting pressure on the whole market. This fact, as well as the increased transparency due to the high distribution rate of Internet access in Germany, has led to a structural change of the entire banking market. Decreasing customer loyalty and simultaneously decreasing margins characterize the new market

situation. Affiliated and smaller banks in particular are not able to handle their cost structure. Hence, a trend towards more centralization and the reduction of the affiliate banking business can be observed. Since 1995 the number of credit institutes registered in Germany decreased by more than 30%. The number of affiliated banks decreased by almost the same percentage [1]. Against this trend, however, the market for direct banking is increasing. An important cornerstone of this success can be attributed to an underlying cost consciousness [2]. Due to these facts, the market for direct banking will attract more competitors and cost reduction; flexibility and high technical standards will be required to meet the challenges presented by this situation. In order to cope with these challenges, a drastic transformation of the business processes as well as organization and managerial structures are necessary. Additionally, the deployment of information and communication technology (ICT) as well as the reengineering of the available information systems becomes inevitable, supporting and automating not only of the inter- but also the intra-enterprise business processes. The use of business components for the (re) design and (re) engineering of information system provides considerable benefits since they “directly model and implement the business logic, rules and constraints that are typical, recurrent and comprehensive notions characterizing a domain or business area” [3, p. 5].

The idea of building individual software systems by combining pre-fabricated *software components* from different vendors to construct unique applications was introduced for the first time by McIlroy in 1968 [4]. The software components of an information system that support directly the activities in an enterprise are usually called *business components*. A business component provides a set of services out of a given business domain through well-defined interfaces and hides its implementation [5]. The principle of modular design that underlies component-based software systems is equally important for the discussion of the technological as well as the economic advantages of component-based software systems. The compositional plug-and-play-like reuse of components might enable software component markets, where different components can be individually combined according to the customers' need. The general advantage of such systems is widely discussed in literature, c.f. [6-10]. Modular systems have been described as the result of a functional decomposition [11], and the conception of modular systems has been thoroughly analyzed by system theory.

Fundamental for this notion of modular systems is an overall domain-engineering concept in order to gain a perfection of component orientation. For domain engineering [12, p. 19-59, 13, p. 159-169], different domain engineering processes are well-known and in use, among others [14-16]. The methods mentioned contribute to different aspects of domain engineering theory; for example, in identifying prominent or distinctive features of a class of systems or in defining characteristics of maintainability and understandability of a family of systems. A domain engineering method which, throughout all stages of development, addresses the domain in which the business component is used, is the Business Component Modeling (BCM) Process introduced by [17]. Due to being explicitly developed for the context of business components, BCM considers additional prerequisites such as reusability,

marketability and self-containment, which are required for different domains. This paper contributes to the validation of the BCM process and specifically to the validation of the BCI-3D method by applying them to the domain of direct banking customer services.

The outline of the rest of the paper is as follows. In section 765 the BCM process will be briefly explained. Since the identification of business components is still a crucial factor, the Business Component Identification (BCI) method [18] and its further development (BCI-3D) [19] is also shortly described. The BCI-3D method is used for the identification of reusable and marketable business components for the direct banking customer services domain. A detailed description of the customer service domain can be found in section 0. The biggest challenge concerning a specification of the requirements for service oriented IT architectures is an overall and detailed analysis of the considered business domain. The result of this domain analysis is either a defined reference model or a business domain specific domain model or a mix of both. In all cases, reference models as well as domain models have to be derived from the business processes of that specific domain and build the basis for the design of component oriented models and the development of the corresponding component-oriented application systems, as described in section 0. Since this paper contributes to the validation of the BCM process and specifically to the validation of the BCI-3D method, the evaluation results and conclusions are given in section 0.

2 Business Component Modeling Process and the Business Component Identification Method

A precondition to component-based development of application systems by using business components is a stable component model. In order to obtain stable business component models, a well-defined identification process is necessary. The basis for the identification of reusable, marketable and self-contained business components is an appropriate and high quality business domain model. Such a model not only serves to satisfy the requirements for a single application system but rather for a family of systems – and therefore for a certain domain. In order to achieve this, we used the Business Component Modeling (BCM) process [17, 18] as shown in Fig. 1.

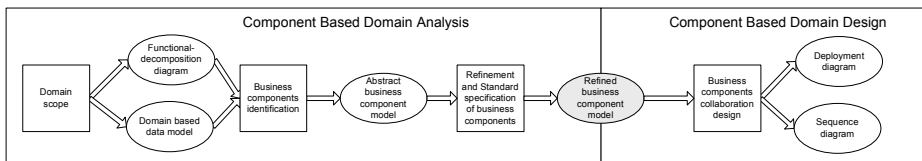


Fig. 1. Business Component Modeling Process [17]

The Business Component Modeling Process can be divided into two major parts, the *Component Based Domain Analysis* and the *Component Based Domain Design*. Rectangles denote sub-phases of the process, and ellipses contain the resulting

diagrams and models of each sub-phase [17]. In this paper we concentrate on the *Domain Scope* and the *Business Components Identification* sub phases and will not describe the *Refinement and Standard Specification* sub-phase. For a detailed description of the whole process we refer to [17]. In the *Domain Scope* sub-phase, a model of the respective domain has to be derived. Therefore the domain of interest is identified, characterized and business processes with their functional tasks are defined. In addition, data is collected to analyze the information objects and their relationships. Possible sources of domain information include existing systems in the domain, domain experts, handbooks, requirements on future systems, market studies, and so on. As a result of the first sub-phase a *functional-decomposition diagram* and a *domain based data model* are generated. Several methodologies and languages can be used to derive the domain model as e.g., DEMO (Design & Engineering Methodology for Organizations) [20-22], ARIS [23], UML (Unified Modeling Language).

The business domain models are fundamental for the next sub-phase in the BCM process, namely the *Business Components Identification* phase. Since the identification of business components is strongly dependent on the quality of the underlying business model, the use of an adequate methodology is absolutely necessary. In order to optimize the process of identifying high quality, reusable and marketable business components the *Three Dimensional Business Components Identification Method* (BCI-3D) has been introduced by Albani et al. [19]. BCI-3D is an extension of the Business Components Identification (BCI) method [18], which has been applied in several case studies such as [24, 25], and which has been improved while considering the evaluation results of the case studies mentioned.

In BCI-3D one can distinguish between three types of relationships necessary for the identification of business components; the relationship between single process steps, the relationship between information objects and the relationship between process steps and information objects. A relationship type distinguishes between subtypes expressing the significance of a relationship. For example, the relationship between single process steps expresses – based on their cardinality constraints – how often a process step is executed within a process and therefore how close two process steps are related to each other in that business domain. The relationship between information objects defines how loosely or tightly the information objects are coupled. In addition, the relationship between process steps and information objects defines whether a corresponding information object is used or created while executing the respective process step. All types of relationship are of great relevance in order to define which information object and process steps belong to which component. The relationships are modeled in the BCI-3D method using a weighted graph. The nodes represent either information objects or process steps and the edges characterize the relationships between the nodes. Weights are used to define the different types and subtypes of relationships and build the basis for assigning nodes and information objects to components. In order to optimize its display, the graph is visualized in a three-dimensional representation having the process steps and information objects arranged in circles and without showing the corresponding weights (see Fig. 5). By satisfying defined metrics such as minimal communication between and maximal

compactness within business components, the BCI-3D method groups process steps and their corresponding information objects with the aim of obtaining an abstract component model in a top-down way. The constraint of providing optimal grouping while minimizing communication means that an optimization problem needs to be solved with a genetic algorithm. The algorithm starts with a predefined solution and improves it by incremental iteration [19]. The starting solution is generated using a greedy graph-partitioning algorithm [26]. For improving the initial solution, the Kernighan and Lin graph-partitioning algorithm [27] has been implemented. The result of applying the BCI -3D method to a defined domain results in a business component model, describing their relationships and provided and/or required services.

To illustrate the domain scope and component identification sub-phases with their resulting diagrams and models, the BCM process is applied to the domain of *direct banking customer services* described in the next sections.

3 The Domain of Direct Banking Customer Service

A sustainable economic growth of direct banks in the German banking sector particularly depends on efficient cost structures as well as cost-effective and quality-conscious customer service processes. Compared to traditional banks, the direct banking sector in Germany is characterized by high rates of customer growth. Due to this fact, new challenges regarding business processes and IT systems arise. The risk of growing internal costs due to inefficiencies of the implemented service processes and service tasks has to be managed at an early stage. Consequently, the business domain *service*, and particularly the domain *customer services*, is in the focus of interest of a direct bank. The business processes in the mentioned business domains are heavily dependent on guidelines and directives. In the example case, these directives are defined and modeled by a company-wide process department. Fig. 4 illustrates an example directive. A directive describes the business tasks, the corresponding responsible organizational units (cost center) and the related resources in a graphical based form. To each of these business process steps, guidelines focuses on legal restrictions, check criteria or documentation needs. The implementation and daily handling of these guidelines and directives are periodically audited and evaluated in an internal review. The sum of all guidelines and directives can be seen as a complete reference model for the domain direct banking customer service. The reference model of one of the leading direct banks in Germany has been used to analyze the domain of customer service.

The actual handling of the customer services is mostly characterized by manual business processes. Information is either entered directly in the supporting IT system or integrated via the web front-end of the customer portal. The main problems are system inflexibility concerning the automation of functionality and integration of new business requirements in the IT system on the one hand and difficulties concerning the full integration of the system landscape, especially the information exchange with the mother bank on the other hand. Due to this fact there are cost and time intensive

inefficiencies in the customer processes as well as supplementary costs of the error handling of the information integration in the system landscape.

In order to model the business functionality of the example domain, functional decomposition diagrams, as used in the ARIS methodology by [23], have been used for describing business tasks and business activities. A functional decomposition diagram splits the main tasks of a business process in four areas: *Task area*, *Function area*, *Sub-function area* and *Elementary function area*. A more detailed splitting of business process tasks is not efficient regarding a general management view. Fig. 2 illustrates for the customer service domain the task area *Service* in a functional decomposition diagram. Due to space limitations, we do not provide the process models which visualize the execution order of the single business tasks and activities.

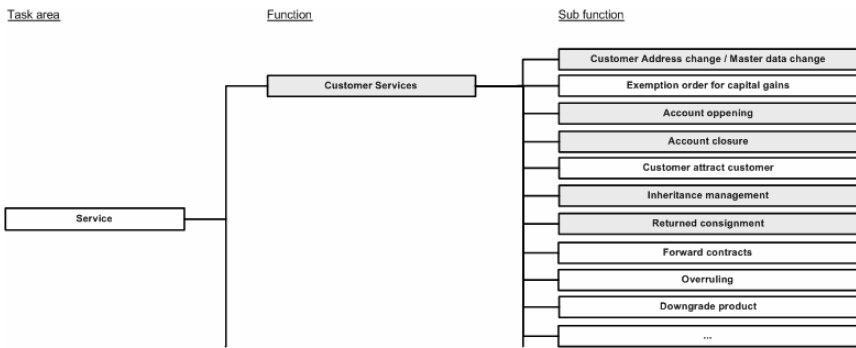


Fig. 2. Functional decomposition diagram for customer services at a direct bank

The task area *Services* provides several functions such as *Customer Services*, *Administration Services* and *Helpdesk Services*. In Fig. 2 only the *Customer Services* function and an extract of its sub-functions are listed. The complete range of sub-functions within the customer services function can be restricted to 27. In the present case, 'relevant' sub-functions were limited to tasks considered in the daily contribution margin calculation. This limitation facilitates the focus on the most commonly used and most expensive tasks and it considers more than 90 % of the whole business activities in the mentioned domain. To get a better understanding of customer services in the direct banking sector, selected sub-functions (highlighted grey in Fig. 2) are described. In this context, the description of tasks is always in the form 'verb - subject - adverb' in order to standardize their description and to avoid ambiguity of the modeling notation.

The sub-function *Customer address change / Master data change* provides tasks like "change name" or "change customer address" for example in case of marriage, divorce or removal. For all these tasks, the differentiation of the communication channel used is important, for example concerning business intelligence requirements, such as calculating the customer service process costs (contribution margin) and controlling the requirements of the different communication channels in the future. A

precondition to change the customer master data via mobile device, Internet or telephone is a valid customer legitimization. Therefore, that sub-function also describes typical direct banking tasks like legitimization of persons by means of the PostIdent proceeding. The German PostIdent proceeding is the standard person legitimization proceeding accomplished by the Deutsche Post AG.

The next logical step after creation and legitimization of a customer is the account opening. Therefore, the corresponding sub-function *Account opening* provides tasks such as “check residential status of customer” in case of an abroad customer address for dispatch. If the customer wants to trade options or financial futures, a clarification of trading risks has to be collected and a risk group has to be calculated. The differentiation of the used communication channels in this sub-function is also an important step.

The last steps in the customer lifecycle are the customer inheritance business task, listed as sub-function *Inheritance management* in the functional decomposition diagram. The first task is the request of an official inheritance confirmation. Afterwards, the bank has to check and delete standing orders for exactly defined cases and criteria. If the customer has a maestro or credit card, these cards have to be locked, and the credit card service provider has to be informed. Additionally, the tax authorities have to be informed about the account balance of the last day before the inheritance. In case of clarified inheritance, the last task is the disclosure of the customer accounts.

The last sub-function we describe is *Returned consignment*. As described above, each of the illustrated sub-functions listed in figure 2 includes several tasks, which are listed in the functional decomposition diagram by means of elementary functions. Due to space limitations we only list and describe the elementary functions of the *Returned consignment* sub-function in more detail.

At first, a returned consignment has to be *checked* and classified. In special cases like PIN numbers or TAN lists, the returned consignment has to be *destroyed* immediately. In all other cases, it has to be verified if the address of the returned consignment matches the most recent customer address for dispatch in the service system (*check customer address in own system*). If the customer address is corresponding and no lock reason is set in the system (*check lock reason returned consignment*), the cover of the new correspondence is *marked with “2”*, and it is then sent to the same customer address for dispatch. If this correspondence returns a second time, the customer address for dispatch is compared to the customer tax address in the host system. In case of a difference between these two customer addresses, the returned consignment is sent a third time to the customer address of tax. At the same time, the customer address for dispatch is deleted in the existing system. In case of a third returned consignment, the customer *account is locked* with the lock reason returned consignment, and the returned consignment is collected in the customer records. If no customer record exists, a new one is to be created. The next task is to *check the customer’s resident status*. If the resident status is “non-resident”, the internal audit is to be informed. If the resident status is “resident”, an *address match* is made with forward orders of the Deutsche Post AG. If there is still no new customer address identified, an *address research* is started. Therefore, the mortality database is scanned. If new information about the customer address cannot be found, the so-called EMA requests are scanned. An EMA request is a request from

other companies like shippers who had similar problems with this customer. If there is no address match in the EMA files, a new EMA entry is made. After 6 months, the customer account is checked for discharge. Therefore, a *supervision charge* is accounted. If a custody account also exists, it will be *disposed* at the same time. Finally, the account balance is *written off* with respect to the defined detailed instructions. The *customer account is deleted*. Finally, all the documents, protocols and official forms have to be *archived*.

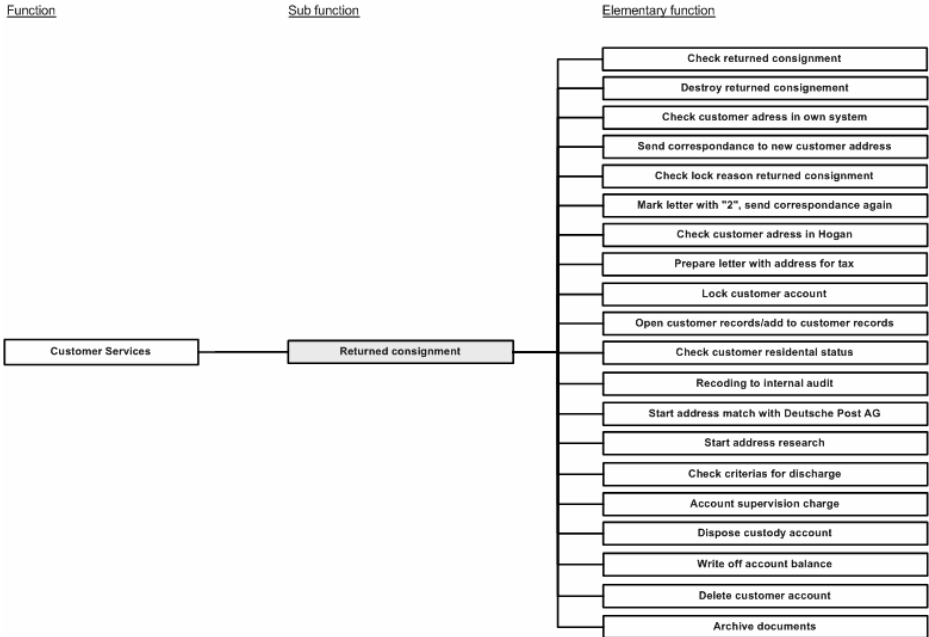


Fig. 3. Functional decomposition diagram of the returned consignment sub-function

A functional decomposition diagram (see Fig. 3) shows all the necessary business tasks with their sub tasks, but does not refer to any information objects, such as customer address, which are used or created while executing those functions. Furthermore there is no chronological order between different tasks. Therefore, the relation between single tasks and information objects has to be modeled in an alternative manner as shown, for example, in Fig. 4.

As mentioned in section 0, there are several methods concerning the modeling of business processes. In the actual case the considered direct bank uses directives and guidelines for the definition of the main tasks, their orders and their dependences. Fig. 4 illustrates such a directive. The graphical modeling is done in the Adonis tool and the provided notation possibilities. Each of these directives and guidelines are provided verbally. The guidelines determine check criteria, supervision periods and

documentation necessities in detail. Every directive additionally indicates the triggering events. In the example under consideration, the triggering event is the physical entry of a returned consignment. The above mentioned *destroy returned consignment* task in case of PIN numbers or TAN lists is not clearly mentioned in the directive, but in the corresponding guidelines. However, the directives give hints about the relation between tasks, depending on the business environment. All available business process models are provided in the local intranet network and show the flag ‘valid’ or ‘invalid’. As consequence it can be assured, that all important business processes are companywide communicated and known.

Having defined all business tasks and the corresponding information objects, the relationships between business tasks, between information objects, and between business tasks and information objects need to be defined in order to use them in the business components identification method. For the purpose of domain analysis, the two task relations *standard* and *optional* were defined. A standard relationship between tasks arises from the business process structure. In Fig. 4 the arrows reflect this relationship. The stringent following of each arrow with regard to the different tracking possibilities describes a standard relationship apart. The starting point of all standard relationships is always the triggering event. Not taking into account the final process course all standard relationships end at the directive-ending node. The second relationship type between different tasks is the optional relationship. In this context, an optional relationship refers to a possible relation between two tasks, clearly depending on the predetermined guidelines. For instance the creation of minutes during a customer conference call is not always mandatory.

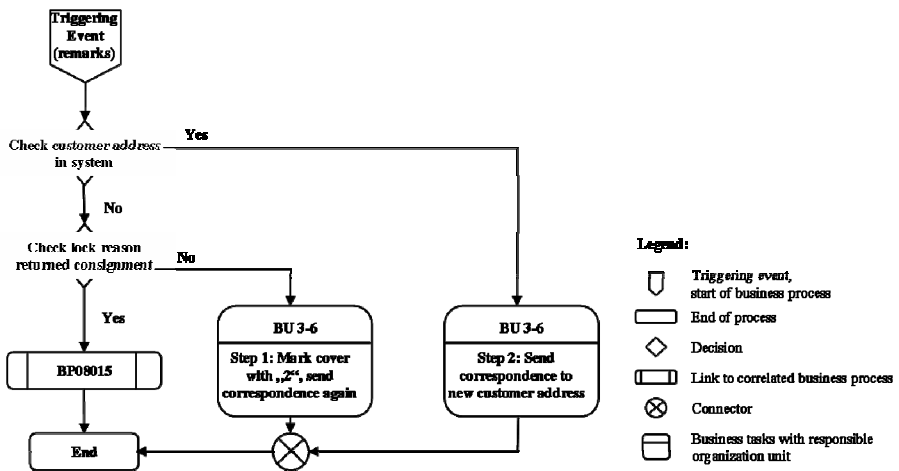


Fig. 4. Business process directive: handling of returned consignment

The authors decided to distinguish between a standard and an optional relationship to have the possibility of different weightings during the business component identification. With these relationships already defined, the communication effort of

the identified business components is minimized due to the above-mentioned algorithms of the BCI-3D method.

In addition to the relationships between tasks, the relationships between task and information objects have to be considered. As already stated, a task can *create* information objects (outgoing information object) or *use* (incoming) information objects. For example, a customer correspondence can only be sent if the customer address for dispatch is known. To execute the task *Sent customer correspondence* the incoming information objects *customer address for dispatch*, *stamp*, *envelope* and *prepared customer correspondence* are required.

Finally, the relationships between information objects have to be provided for efficient business component identification. Three different types of relationships are defined in this case: *related-to*, *part-of* and *state-of*. In general, a *related-to* relationship is defined between entities like customer and contract. A customer has at least one or more contracts whereas a contract is attributed exactly to one customer. The *relationship part-of* signals that an information object is part of another information object. For example, the information objects *customer addresses for dispatch* and *customer tax address* are part of the information object *customer master address data*. Finally the relationship *state-of* signals a change in the condition of an information object. For instance, a customer contract runs through the conditions *customer request* and *offered contract* before becoming the status *customer contract*.

All the mentioned relationships belong to the set of relationships provided and recommended by the BCI-3D method. In order to gain all relevant information for modeling the supporting information system by means of business components, ten representative customer service processes were modeled in the described proceeding, including business tasks, information objects and all their relationships as mentioned above. All the considered services processes are highlighted in Fig. 3. Applying the BCI-3D to the services mentioned resulted in a business component model as illustrated in section 0.

4 Business Component Model of Customer Service Processes

The modeling of the above mentioned customer service processes leads to 288 different information objects and 159 business process steps (BPS). Concerning the information objects it was not distinguished between paper based and digital information objects. In order to use the business component identification method for identifying reusable and marketable business components, all relationships need to be weighted. The reason therefore is to define how strong objects (process step or information objects) are related in order to ensure their grouping within one and the same business component. With this approach, individual business process structures of the direct bank case are analyzed in detail. The weights result from feedback received while using the BCI-3D method and from the detailed analysis of the actual business domain. In the actual case, the weighted relationships were defined as follows:

- a. Weights for relationships between information objects:
 - i. related-to: 100
 - ii. part-of: 1000
 - iii. state-of: 1000
- b. Weights for relationships between business tasks:
 - i. standard: 1000
 - ii. optional: 100
- c. Weights for relationships between business tasks and information objects:
 - i. create: 1000
 - ii. use: 100

As a result, the main focus of the business component identification will be on minimal communication efforts between different *standard* tasks, ensuring loosely coupling of business components, and on tight cohesion, ensuring optimal grouping of related business functionality. In addition, each information object, which is created by a specific business task, is placed within the same business component as the business tasks itself. Fig. 5 illustrates the identified component model.

Altogether, 21 business components were identified. Fig. 5 illustrates the single component “*service component*” on the left side and the complete component model on the right side. All BPS are placed ‘on the ground’ of the component model, whereas the information objects are situated above, ‘on top’ of the component model. The mentioned relationships between BPS and information objects are represented by connectors. It can be seen, that the main communication of the single component is inside the component. The complete component model on the right side strengthens this statement. The communication efforts between the single components are minimized.

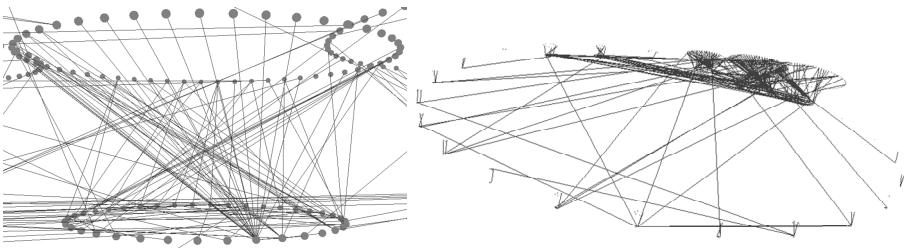


Fig. 5. Component model for customer services

The main components identified are: *archive component*, *service component* and *master data component*. The biggest component on the right side obviously shows more external rather than internal communication. This indicates that the information objects of this component are created by business tasks located in other components. This observation could be seen as a contradiction of the main focus, which is less communication between the business components within the aforementioned relationships. However, a closer examination of the functionality provided by this business component (i.e. the *archive component*), provides us with

a clearer understanding of this interaction. This is to say, when protocols, forms and customer requests are demanded, they are stored and offered as service in the archive component. That way, the identified communication structure is efficient. Information objects, which are needed for executing a task within a component, are created in the service component, not in the archiving component. The other two sizeable components provide the main data concerning customer master data or the account master data with the main tasks fitting customer creation and account opening.

With the structure of this component model, the characteristics of the customer services domain become obvious. There are independent customer requests for services. In general, a customer requests banking products and services step by step. According to this perspective, the identified business components generally contain one service with all the relevant tasks to fulfill each of these steps. Furthermore, where possible these components are defined as discreet and independent. Fig. 6 illustrates an extract of the component model for the customer process *customer attracts customer*, following the UML 2.0 notation [28] presenting the components with their required and provided services.

In reality, the customer service process *customer attracts customer* is mainly manually implemented. As a result, this process provides a lot of effort and produces costs and errors. Based on the domain analysis, the BCI-3D tool suggests four components for handling this process. The business component *customer attracts customer* provides all tasks for recording the service process. To check the correctness of a request to this component, the service *get sales prospect information* is necessary to import all related sales prospect information about the prospective customer and the new customer. If the request fits the business rules, the service *offer confirmation customer attracts customer* is provided in the component model. At the same time, a letter of thanks to the attracting customer is prepared. Therefore, the actual customer address for dispatch is provided in the component model by means of the service *get actual customer address*.

The second business component deals with customer requests to change the customer gift, which will be sent to new customers. Therefore the actual shipping list of the customer gift has to be requested in the component model via *get shipping list presents*. If the customer gift is not yet foreseen in the actual shipping list, a change of the customer gift is still possible and a confirmation of the new gift is provided in the component model by the service *offer confirmation customer attracts customer*. At the same time, this component provides the general service of checking a customer gift in the actual shipping list via *offer present not in actual shipping list*.

This service is necessary to fulfill the tasks of the third component. This component prepares customer claims. With the mentioned service of the second component, it is possible to check if an unsent customer gift is foreseen in the actual shipping list. The business process determines the validity of a shipping list up-to one week. Therefore, the following shipping list should also be checked. The shipping list is demanded by the fourth component via *get shipping list present*. If the customer gift is not included in any shipping list, the service *offer error present* is provided in the component model to business component 1 to restart the recording process. If there is a

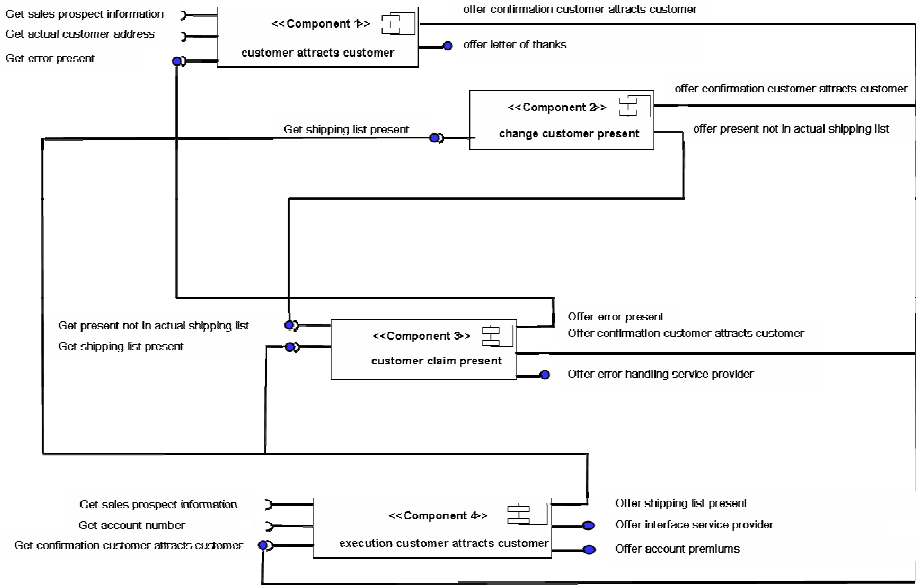


Fig. 6. Component model – business components customer attracts customer

problem with the quality of the customer present, the service *offer error handling service provider* sends the necessary information directly to the service provider.

The fourth business component presents the main interface to the service provider. This component maintains the shipping lists. If a customer present is chosen in any of the other components, this information is included via the service *get confirmation customer attracts customer*. Consequently, this component also offers the actual shipping lists to the component model via *offer shipping list presents*. Each week the customer gifts are demanded by the service provider via the service *offer interface service provider*. If a customer wishes to get a bonus instead of a gift, this component also prepares the accounting of the bonus. Therefore, the actual customer account number has to be demanded in the component model via *get account number*.

The extract of the business components *customer attracts customer* gives an initial insight into automating business processes. All tasks except the customer claim can be implemented and automated, at least for customer requests via the Internet. The identified components provide not only services to fulfill the customer service process, but these services easily can be reused by other components in other business domains. With such an IT-architecture, a focus on the important customer request – the claiming requests – is possible without losing quality for the other task. In contrast, the automation of the mentioned tasks provides faulty insertion and correction in the actual systems. A customer can check immediately and independently whether a change of his gift is still possible. Additionally he could track the proceeding of this service process.

5 Evaluation and Conclusions

The initial purpose for writing this paper was a six-month project at one of the five leading direct banks in Germany. The challenge was to reduce costs by automating data warehouse processes and to increase the quality of customer services at the same time. Therefore the domain *customer services* needed to be analyzed in detail. In order to use a methodological approach, the business components modeling process (BCM) was applied parallel to the normal project work, focusing mainly on the first two sub-phases, the domain scope and the business component identification. The methodology was used for 10 of the mentioned 27 customer services with the objective of developing a service-oriented approach for an alternative implementation of the supporting IT-landscape.

Advantages of the use of the BCM methodology

Applying the domain scope sub-phase to the example domain meant to analyze all business tasks and the relevant information objects in detail, allowing a more precise definition of the process costs. New service processes, which led to further margin calculation costs, were identified. Additionally, an overall service process documentation was created which led to a deep acceptance and support of the general proceeding of the BCM methodology in the business unit.

With the business components identification method, based on the resulted domain models, an automated approach was provided for the identification of business components. The metrics defined in this paper – being minimal communication between and maximum compactness of business components – are the basic metrics for the component-based development of supporting information systems. Based on the well-defined business task dependencies, business functions could be clustered within business components. The identified components have a potential of reuse due to the provided services. In consequence, redundant activities in the business process as well as technical redundancy in the data warehouse functionalities were identified. This fact led to direct cost reductions in the services processes and to indirect cost reductions concerning a reduced error rate and reduced maintaining costs in the data warehouse.

Furthermore, a huge potential, especially for the market of direct banks, arises on the basis of process automation. Thereby a concentration on core competencies can be achieved without losing quality in the existing customer services. In the example service process *customer attracts customer* 80% of the process costs occurred for activities with automating potential, only 20% for activities concerning the customer claim. The result of the BCI methodology is therefore used to calculate the benefit of a service oriented architecture on the base of the customer service processes due to process automation potential. In consequence, an individual business case for the implementation of a service-oriented architecture can be calculated based on a methodical approach. In case of a positive business case, the identified component model represents the complete conception at the same time.

Difficulties by the usage of the BCM methodology

Since the identification of business components is strongly dependent on the quality of the underlying business domain models, a strong focus was set on modeling the example domain. During that phase, difficulties occurred due to the way of modeling,

especially with regard to the allocation of information objects. In most cases, the information objects could only be gained from the guidelines that were delivered with the specific processes. Therefore, depending on individual policies one process-guide could be described more precisely than another. Due to these reasons, a consistent level of abstraction for modeling meaningful business processes was not given. By contrast, the relationships between the functions were documented adequately within the project case. Accordingly, when analyzing the component model, some components contain only highly abstract information objects, as for example *scoring product* that represents the score value of an individual product. Polishing these information objects, for example to include scoring per product, would improve the correlation of these highly aggregated information objects.

The examination of ten customer services showed, that guidelines and directives are not sufficient to exploit the complete power of the BCM method. In consequence, use of the BCI-3D tool only makes sense to a certain level of abstraction. Additionally, by analyzing more than only ten service processes even better reusable business components would result.

Hints for further developments for the BCI tool

Concerning direct banking requirements, security aspects on data have to be considered. Actually it is not possible to integrate new, sector specific dimensions like *data security* to the BCI-3D method. But these aspects influence in an important way the identification of business components. If a role concept could be considered during the component identification a big manual task of post processing could be avoided.

Another actual weakness of the BCI-3D tool implementing the BCI-3D method is the lack of interfaces for example concerning data dictionary or metadata tools. In most direct banks, the data models are maintained. If the reuse of individual direct bank data domains would be possible, the quality of the identified business components would be higher. Another interesting point is the integration of automatic business logic checks via the data load in BCI-3D. If the tool would provide a meta model to rebuild business dependencies, logical errors in the domain analysis could be detected.

Problems regarding the use of the BCI-3D tool arise from filling the flat files that constitute the basis for the optimization method, which results in enormous manual work while mapping the information gained in the domain model to the representation of the BCI-3D tool. Certainly, the most comfortable way of optimizing that step would be the integration of the tools for modeling the business domain and the BCI-3D tool.

References

1. Bundesverband deutscher Banken <http://www.bankenverband.de/index.asp?channel=168247&art=769>. 2005.
2. Online Banking: Der Zuwachs ist ungebrochen. <http://www.bankenverband.de/channel/133810/art/1535/index.html>. 2005.

3. Barbier, F. and Atkinson, C., Business Components, in Business Component-Based Software Engineering, F. Barbier, Editor. 2003, Kluwer Academic Publishers Group. p. 1-26.
4. McIlroy, M.D. Mass Produced Software Components. In Software Engineering: Report on a Conference by the NATO Science Committee. 1968. Brussels: NATO Scientific Affairs Division.
5. Fellner, K. and Turowski, K. Classification Framework for Business Components. In Proceedings of the 33rd Annual Hawaii International Conference On System Sciences. 2000. Maui, Hawaii: IEEE.
6. Baldwin, C.Y. and Clark, K., Managing in an age of modularity. Harvard Business Review, 1997. 75 5: p. 84-93.
7. Baldwin, C.Y. and Clark, K., Design Rules: The Power of Modularity. 2000, London: MIT Press, Cambridge (Mass.).
8. Sanchez, R., Strategic flexibility in product competition. Strategic Management Journal 16, 1995: p. 135-159.
9. Sanchez, R. and Mahoney, J.T., Modularity, flexibility and knowledge management in product and organization design. Strategic management Journal, 1996. 17: p. 63-76.
10. Schilling, M.A., Toward a general modular systems theory and its applications to interfirm product modularity. Academy of Management Review, 2000. 25: p. 312-334.
11. Ulrich, K.T., The role of product architecture in the manufacturing firm. Research Policy, 1995. 24: p. 419-440.
12. Czarnecki, K. and Eisenecker, U.W., Generative Programming: Methods, Tools, and Applications. 2000, Boston: Addison-Wesley.
13. Sametinger, J., Software engineering with reusable components. 1997, Berlin; New York: Springer. xvi, 272 p.
14. Kang, K., et al., Feature-Oriented Domain Analysis (FODA) Feasibility Study. 1990, Carnegie-Mellon University: Pittsburgh, PA.
15. Simos, M., et al., Organization Domain Modeling (ODM) Guidebook. 2.0 ed. Informal Technical Report for STARS STARS-VC-A025/001/00. 1996.
16. D'Souza, D.F. and Wills, A.C., Objects, Components, and Frameworks with UML: The Catalysis Approach. 1999, Reading: Addison-Wesley.
17. Albani, A., et al. Domain Based Identification and Modelling of Business Component Applications. In 7th East-European Conference on Advances in Databases and Informations Systems (ADBIS-03), LNCS 2798. 2003. Dresden, Deutschland: Springer Verlag.
18. Albani, A., Dietz, J.L.G., and Zaha, J.M. Identifying Business Components on the basis of an Enterprise Ontology. In Interop-Esa 2005 - First International Conference on Interoperability of Enterprise Software and Applications. 2005. Geneva, Switzerland.
19. Albani, A. and Dietz, J.L.G. The benefit of enterprise ontology in identifying business components. In IFIP World Computing Conference. 2006. Santiago de Chile.
20. Dietz, J.L.G., The Atoms, Molecules and Fibers of Organizations. Data and Knowledge Engineering, 2003. 47: p. 301-325.
21. Dietz, J.L.G. Generic recurrent patterns in business processes. In Business Process Management, LNCS 2687. 2003: Springer Verlag.
22. van Reijswoud, V.E., Mulder, J.B.F., and Dietz, J.L.G., Speech Act Based Business Process and Information Modeling with DEMO. Information Systems Journal, 1999.
23. Scheer, A.-W., ARIS - Business Process Modeling. 2 ed. 1999, Berlin: Springer.

24. Selk, B., Klöckner, K., and Albani, A. Enabling interoperability of networked enterprises through an integrative information system architecture for CRM and SCM. In *International Workshop on Enterprise and Networked Enterprises Interoperability (ENEI 2005)*. 2005. Nancy, France.
25. Selk, B., et al. Experience Report: Appropriateness of the BCI-Method for Identifying Business Components in large-scale Information Systems. In *Conference on Component-Oriented Enterprise Applications (COEA 2005)* in conjunction with the Net.Objectdays. 2005. Erfurt, Germany.
26. Jungnickel, D., The Greedy Algorithm, in *Graphs, Networks and Algorithms*, D. Jungnickel, Editor. 2005, Springer: Berlin. p. 123-146.
27. Kernighan, B.W. and Lin, S., An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal*, 1970. 49: p. 291-307.
28. OMG, *OMG Unified Modelling Language, Version 2.0*, in *Secondary OMG Unified Modelling Language, Version 2.0*, Secondary OMG, Editor. 2003: Place Published. p. Pages.http://www.omg.org/technology/documents/modeling_spec_catalog.htm#UML.

An Event-Trigger-Rule Model for Supporting Collaborative Knowledge Sharing Among Distributed Organizations

Minsoo Lee², Stanley Y.W. Su¹, and Herman Lam¹

¹ Database Systems Research and Development Center
University of Florida, Gainesville, Florida 32611, U.S.A.
Tel: 352-392-2693, 352-392-2689; Fax: 352-392-3238

{su, hlam}@cise.ufl.edu

² Dept of Computer Science and Engineering, Ewha Womans University,
11-1 Daehyun-Dong, Seodaemoon-Ku, Seoul, 120-750, Korea
Tel.: +82-2-3277-3401; Fax: +82-2-3277-2306
mlee@ewha.ac.kr

Abstract. The Internet has become the major platform for future inter-organizational knowledge-based applications. There is a need for knowledge modeling and processing techniques to perform event management and rule processing in such a distributed environment. We present an Event-Trigger-Rule (ETR) model, which differs from the conventional ECA rule model in that events and rules can be defined and managed independently by different people/organizations at different sites. Triggers are specifications that link distributed events with potentially complex structures of distributed rules to capture semantically rich and useful knowledge. Triggers can also be specified by different people/organizations in a distributed environment. Based on the ETR model, we have implemented an ETR Server that can be installed at multiple sites over the Internet. The ETR Server provides platform independence, extensibility, processing of event histories and rule structures, dynamic rule change at run-time, and Web-based GUI tools. The ETR Model and the implemented ETR Server can support various inter-organizational collaborative knowledge-based applications such as a Web-based negotiation system, supply chains, dynamic workflow management system, Knowledge Networks, and transnational information system.

1 Introduction

The Internet has become the major platform for future inter-organizational knowledge-based applications. Such applications will also need to become active, as more machine processing rather than human interactions are required on the Internet. Therefore, these inter-organizational knowledge-based applications need knowledge modeling and processing techniques to perform event management and rule processing in a distributed environment. By using rules, a high level specification of knowledge can be used to enhance the system with active capabilities. In an active system, rules are automatically executed when a particular event of interest occurs

and some data condition is satisfied. The rule system can automatically carry out security and integrity constraint checking, business regulation/policy enforcement, alerting, automatic backup and recovery, etc. to eliminate error prone and tedious tasks that required human interventions in passive systems.

Rules used in current knowledge-based systems are specified in a variety of forms. One form of rule specifications, which is very commonly used in active database systems, is the Event-Condition-Action (ECA) rule specification. An ECA rule consists of three parts: the event, condition and action part. The semantics of an ECA rule is that when an event occurs, the condition is checked. If the condition evaluates to true, the action is executed. Otherwise, the action is not executed. In an ECA-rule-based system, the E, C and A specifications form a single rule and is usually managed and processed by a centralized system.

In this paper, we present an Event-Trigger-Rule (ETR) model, which differs from the conventional ECA rule model in that events and rules can be defined and managed independently by different people/organizations at different sites. Triggers are specifications that link distributed events with potentially complex structures of distributed rules to capture semantically rich and useful knowledge. They again can be specified by different people/organizations in a distributed environment. Based on the ETR model, we have implemented an ETR Server that can be installed at multiple sites over the Internet. The ETR Model and implemented ETR Server enables various inter-organizational knowledge-based applications such as a Web-based Negotiation Server[1], a supply chain scenario[2], a dynamic workflow management system[3], a Knowledge Network infrastructure[4], and a transnational information system [5]. We provide a detail description of our ETR Model and ETR Server implementation and suggest that our ETR Model and ETR Server can be used as a basic modeling tool and infrastructure for various advanced inter-organizational systems.

The organization of the paper is as follows. Section 2 discusses the related research. Section 3 gives a detail explanation of our ETR model and section 4 gives an overview of our ETR Server implementation. Section 5 provides some examples of using the ETR Model and ETR Server for inter-organizational systems that we have developed. And section 6 finally gives the conclusion.

2 Related Research

The concept of rules was originally introduced in the research areas of artificial intelligence and expert systems. One popular form of rule specification, the Event-Condition-Action (ECA) rule, has been used in the database management area to create a new category of databases, namely, active databases [6]. As we pointed out in the introduction section, these rules are used in centralized active database systems. Our Event-Trigger-Rule model, although shares some components of rule specification with the ECA model, is designed for distributed knowledge sharing and application development and has several added features.

WebRules [7] is a framework to use rules to integrate servers on the Internet. The WebRules server has a set of built-in events that can notify remote systems, and has a library of system calls that can be used in a rule to connect Web servers. However, it

does not include advanced concepts required for knowledge sharing such as event and rule publishing or event filtering.

Several content-based event notification architectures have been recently proposed to provide an abstraction of the communication infrastructure on the Internet. These architectures focus on providing a scalable architecture for event delivery, as well as a mechanism to selectively subscribe to information. Siena [8], Kyrex [9] and CORBA Notification Service [10] are such proposed event architectures. In [11], the authors outline the concepts of event description, subscription, and notification. One can subscribe either to a single event or an event pattern. The latter is similar to the event structure (or composite event) supported by our ETR Server [4]. A key feature of [12] is the composition of high-level events from low-level ones.

3 Event-Trigger-Rule Model

The ETR model provides the basic modeling constructs such as classes, objects, attributes/properties and methods just like most existing object models. Additionally, it provides three knowledge modeling constructs: event, rule and trigger. These are shown in Figure 1.

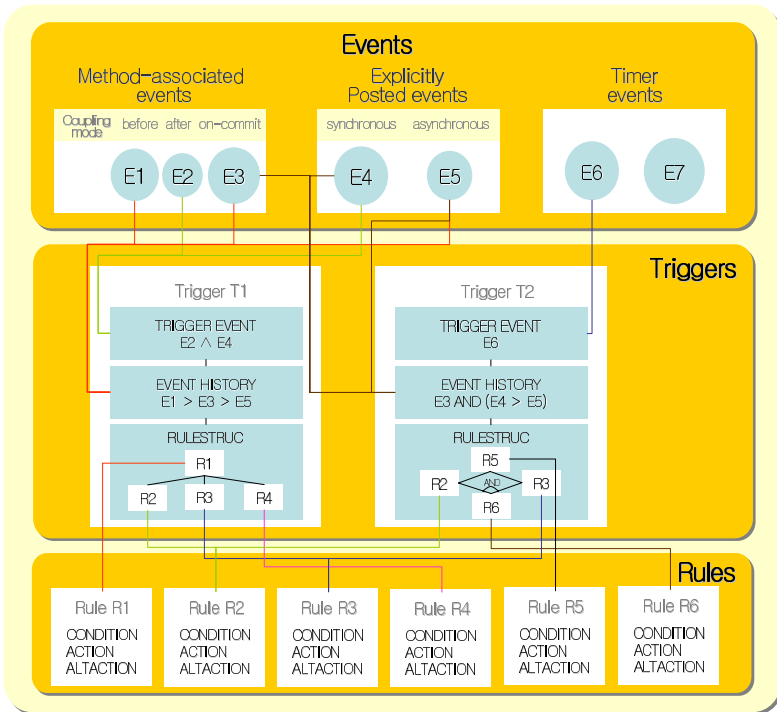


Fig. 1. The ETR (Event-Trigger-Rule) Model

Events model the occurrences of interest, and rules contain specific knowledge conditions and actions. Triggers provide a way to model the complex relationships between the events and rules in a very flexible way. The details of each of these knowledge components are explained below.

3.1 Event

An event is an occurrence of anything of interest to people or software systems. It can be the reading or updating of a data field or record, or a failure signal issued by the software that controls a disk, or the signal of a timer, etc. Events can be grossly categorized into three types: method-associated events, explicitly posted events, and timer events.

Method-associated events. When a specific method executes, the event associated with the method is raised. The raising of an event can be done either before the method, after the method, or at the commit time of a transaction, which contains the method execution. These different times of the raising of events with respect to the method execution time are called coupling modes. The three coupling modes just illustrated are called before, after, and on-commit. These three coupling modes are synchronous. When a synchronous event is posted, the execution of the event posting method/program is suspended until a response is returned from the rule processor. If an event is posted asynchronously, the event posting method/program does not wait for the response but continues its execution. The decoupled mode (the fourth mode) allows an event to be posted asynchronously.

Explicitly posted events. An explicitly posted event can be raised independent of any method execution. That is, the event is not tied to any specific method and can be raised in the body of any desired method via a 'PostSynchEvent' or 'PostAsynchEvent' call. An explicitly posted event can be posted synchronously or asynchronously.

Timer events. A timer event is an event that is related to some predefined time of interest. It is raised when the predefined time of interest has come. The timer event is asynchronous by nature.

Figure 2 shows an example for specifying a method-associated event 'update_quantity_event'. The 'update_quantity_event' will be posted before the activation of the 'UpdateQuantity' method of the 'InventoryManager' class. The event will carry the quantity and item parameters given to the UpdateQuantity method.

IN	InventoryManager
EVENT	update_quantity_event (String item, int quantity)
TYPE	METHOD
COUPLING_MODE	BEFORE
OPERATION	UpdateQuantity (String item, int quantity)

Fig. 2. Specification of a method-associated event : update_quantity_event

3.2 Rule

A rule is a high-level declarative specification of a granule of executable code that is related to an event or events. A rule is composed of a condition, action, and alternate action clause. When an event is posted, the rules associated with the event are triggered for processing. When a rule is triggered, the condition clause of the rule is first evaluated. If the condition is true, the statements in the action clause are executed. Otherwise, the statements in the alternate action clause are executed. A rule has an interface that specifies what parameters are used in the rule body (i.e. condition, action, alternate action). The actual values of these parameters are provided by the event at run time. The syntax of a rule is shown in Figure 3.

RULE	rule_name (parameter list)
[RETURNS	return_type]
[DESCRIPTION	description_text]
[TYPE	DYNAMIC/STATIC]
[STATE	ACTIVE/SUSPENDED]
[RULEVAR	rule variable declarations]
[CONDITION	guarded expression]
[ACTION	operation block]
[ALTACTION	operation block]
[EXCEPTION	exception & exception handler block]

Fig. 3. Syntax of the rule language

In the rule syntax, clauses that are surrounded by brackets are optional. A rule can return a value whose type is indicated by the RETURNS clause. The TYPE clause indicates whether or not the rule is allowed to be modified after the initial definition. A DYNAMIC rule indicates that the rule may be changed at run-time, whereas a STATIC rule means that the rule is less likely changed. This information is used for optimizing the performance of the rule by generating proper forms of rule codes. The STATE clause indicates whether the rule will be initially active or suspended after definition. The RULEVAR clause has the declaration of the variables that are used in the rule body. There are four types of rule variables: temporary type, persistent type, existing type, and customizable type. The temporary type is used like a local variable in the rule and the persistent type will persist values of the variable. The existing type enables referencing of remote server objects and the customizable type enables values to be customized for different execution contexts. The CONDITION clause is composed of a guarded expression. A guarded expression has two parts: a guard part and a condition expression part. The guard part is composed of a sequence of expressions. If any expression in the guard evaluates to false, the rule is skipped. Otherwise, the condition expression part is evaluated. The guard part is provided to screen out cases where the rule must be skipped such as error situations or invalid data values, etc. Depending on the result of the condition expression evaluation, the ACTION clause or the ALTACTION clause is executed. ACTION and ALTACTION clauses are composed of statements to be executed such as method calls or assignment statements, statements that post events, etc. During the execution of a rule, an

exception may occur and this is handled by the EXCEPTION clause where the exception type is paired with an exception handler.

An example of a simple rule specification is shown in Figure 4. The rule monitors the change of the “quantity” value. The rule only monitors the item ‘Part225’ and will only proceed if the quantity decreases. This is specified via a guarded expression in the CONDITION. If the available quantity of the item in stock decreases more than 500, the item needs to be ordered from the retailer at an amount of 110% of the quantity change. Otherwise, it only orders the amount that decreased. The RULEVAR supports various types of variables that can make the rule body much simpler. The 'existing' keyword in the RULEVAR indicates that the variable 'retailer' is to be bound with an existing Retailer object in the distributed environment. The Retailer object should have the ID of RETAILER1223.

RULE	check_quantity_change (int quantity_change, String item)
DESCRIPTION	if the quantity is decreased by more than 500, order item from Retailer by 110% of quantity change
RULEVAR	existing Retailer retailer("RETAILER1223"); int order_amount;
CONDITION	[item='Part225' & quantity_change < 0] quantity_change < -500
ACTION	order_amount = 1.1 * (-quantity_change); retailer.Order(order_amount, item);
ALTACTION	order_amount = -quantity_change; retailer.Order(order_amount, item);

Fig. 4. Example rule

3.3 Trigger

Now that the events and rules are specified, the trigger which relates the events and rules can be specified. A trigger basically specifies which events can trigger which rules. It also can support composite events and does the parameter mapping between the event parameters and rule parameters. Another important function of the trigger is the capability to specify the various rule execution sequences. A trigger has the syntax shown below in Figure 5.

TRIGGER	trigger name (trigger parameters)
TRIGGEREVENT	events connected by OR
[EVENTHISTORY	event expression]
RULESTRUC	structure of rules
[RETURNS	return_type : rule_in_rulestruct]

Fig. 5. Syntax of trigger

The trigger has five clauses. The clauses that are surrounded by brackets are optional. The TRIGGER clause specifies the name of a trigger and the trigger parameters. The trigger parameters are used to bridge between the event parameters and the rule

parameters. The TRIGGEREVENT clause lists the events that can trigger the set of rules in the RULESTRUC clause. Several events can be OR-ed (i.e., connected with a disjunctive operator), which means that the occurrence of any one of the events can trigger the rules. The EVENTHISTORY supports checking of past event occurrences that can form composite events. The RULESTRUC clause specifies the set of rules to be executed and also in what order the rules should be executed. Any rule execution sequence is typically composed of three kinds of elements. These elements are sequential rule execution, parallel rule execution, and synchronization points. The RETURNS clause is optional, and is used when the trigger needs to return a value in the case when a synchronous event has triggered the rules. The return type as well as the specific rule that should provide the return value is specified in this clause.

Figure 6 shows a complex trigger that contains an expression within the TRIGGEREVENT clause and EVENTHISTORY clause. The EVENTHISTORY checks if there has been an occurrence of event E4 happening before event E5. The RULESTRUC clause contains several rules to be executed in a linear fashion. The parameter mappings from event parameters to trigger parameters, and again trigger parameters to rule parameters are shown.

```

TRIGGER          sample_trigger ( int v1, int v2, classX v3 )
TRIGGEREVENT    E1( int v1, int v2, classY t1, classX v3 ) OR
                E2( int v2, classX v3, int v1 )
EVENTHISTORY    E4 > E5
RULESTRUC       R1(v1,v2) > R2(v3,v1) > R3(v1,v2,v3)

```

Fig. 6. Example complex trigger

4 ETR Server Design and Implementation

We have implemented an ETR Server that can process the triggers and rules upon receiving an event. Figure 7 illustrates the architecture of the ETR Server. The ETR Server has an extensible interface which can support various communication infrastructures such as RMI, CORBA, Vitria Communicator and a Call Level Interface (CLI) to be used by the administrator of the ETR Server. The Rule Object Manager is the entry point to the core ETR Server. There exist two hash tables: event hash table and the trigger hash table. The event hash table stores the event to trigger hashing information. By hashing with the event name, the corresponding trigger can be identified. The trigger hash table stores the mapping from the trigger to its triggering events. The Rule Scheduler can perform the scheduling of multiple rules. The rules are executed by threads. The Event History Processor evaluates the event history part of a trigger. The Rule Group Coordinator manages the rule group information.

The ETR Server has the following features.

- **Rule Scheduling:** The scheduling algorithm uses several data structures that mainly store information about the predecessors and successors of each rule and the event parameter mappings to the rules. The scheduling algorithm is as follows. Whenever a rule is finished, each of its successor rules are checked and can be started if the predecessor rule completion requirements are met.

- **Event history processing:** Event history processing is performed by the Event History Processor (EHP). The EHP is integrated with the ETR Server to evaluate complex event histories very efficiently. The complex event histories are represented as event graphs and are stored within the EHP. When the ETR Server receives an event, the EHP will update its event graph to pre-compute the result of the event history expression. This result is immediately provided when asked for.

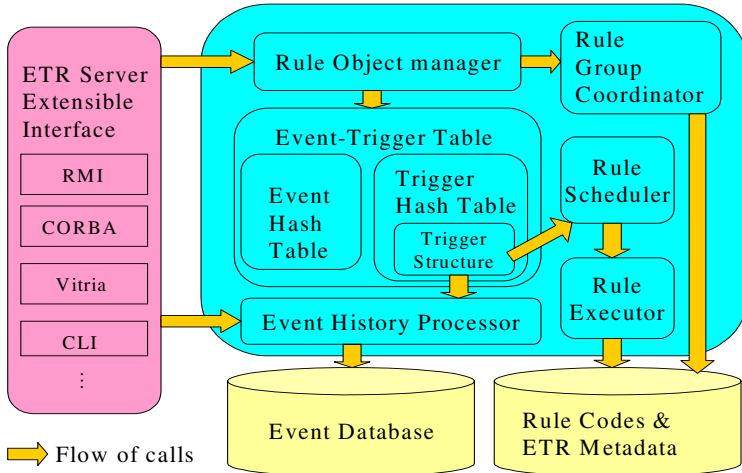


Fig. 7. Architecture of the ETR Server

- **Dynamic rule change:** The ETR Server supports dynamic rule change where rule instances running in the previously defined form can continuously be executed while instances of the newly modified rule can be created and executed.
- **Rule Variables:** The rule variables in the rule are translated into special segments of Java code. The temporary rule variables are translated into ordinary Java variables. The persistent rule variables are translated into a data retrieving and storing command from the persistent store. The existing rule variables are translated into segments of code that are relevant to the type of distributed object infrastructures commands. The customizable rule variable allows the rule to be customized based on several different people or organization's needs. Each person or organization will specify the desired value for the customizable rule variable, and the values are stored into a hash table dedicated to the customizable rule variables. When the rule is being executed, the person or the organization that invoked the rule is identified and the value can be pulled out of the hash table.
- **Rule group:** Rules are grouped and can be activated or deactivated as a group. The rule groups are managed by a Rule Group Manager within the ETR Server. The Rule Group Manager stores information about which rule is included in which rule group. It also stores information about which group is currently active or suspended.

5 ETR Technology in Collaborative Inter-organizational Systems

We have applied the ETR Technology to various applications and suggest that it is suitable as a basic infrastructure for building inter-organizational systems. The following subsections give an idea on how the ETR Server can be used for applications such as a Negotiation Server for automated negotiation and Knowledge Networks for distributed knowledge sharing.

5.1 Automated Negotiation System

Automated negotiation is an essential function in the support of collaborative e-business over the Internet. We have developed an Automated Negotiation System to perform automated negotiation based on the rules stored in an ETR Server[1]. Figure 8 shows the architecture of the Negotiation System. During the build-time, the Policy Maker and the Negotiation Expert of a business enterprise specify their business goals, policies, strategies, decision-action rules by using the GUIs provided by a Knowledge Acquisition Processor. Some other negotiation knowledge, such as mini-world information, is imported to a Knowledge Repository by using the APIs provided by a Knowledge Manager. Negotiation policy rules, strategic rules, and decision-action rules are stored in the Rule Base of the ETR Server.

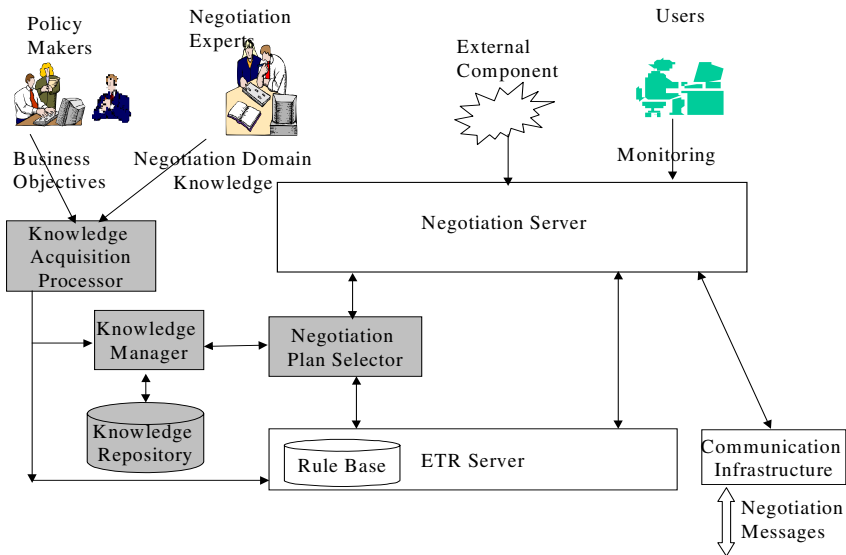


Fig. 8. System Architecture of Automated Negotiation System

After the build-time activities for each transaction are completed, activities for each session begin. A Negotiation Server is ready to generate negotiation messages or process incoming messages. It consults with the ETR Server to make a decision at each transition of the negotiation protocol. The ETR Server makes the decision based

on the selected decision-action rules, and sends the decision back to the Negotiation Server. Once the decision is made available, the Negotiation Server generates an appropriate negotiation message and sends it to the counterpart's Negotiation Server through the communication infrastructure. The process continues until either a mutual agreement is reached or one side unilaterally terminates the transaction.

5.2 Knowledge Networks

As the Web has emerged as the new infrastructure for the future, adding knowledge into the Web servers becomes necessary. The knowledge can make the Web servers more active, collaborative, and intelligent. This leads to the idea of incorporating the ETR Server into a package that is an add-on to the current Web servers. The ETR Server can execute the knowledge rules that are installed by various users on the Internet. These knowledge rules are interconnected by events that are posted among the Web servers. This infrastructure, called the Knowledge Network [4], allows the publishers and subscribers of events to contribute their knowledge into the Web.

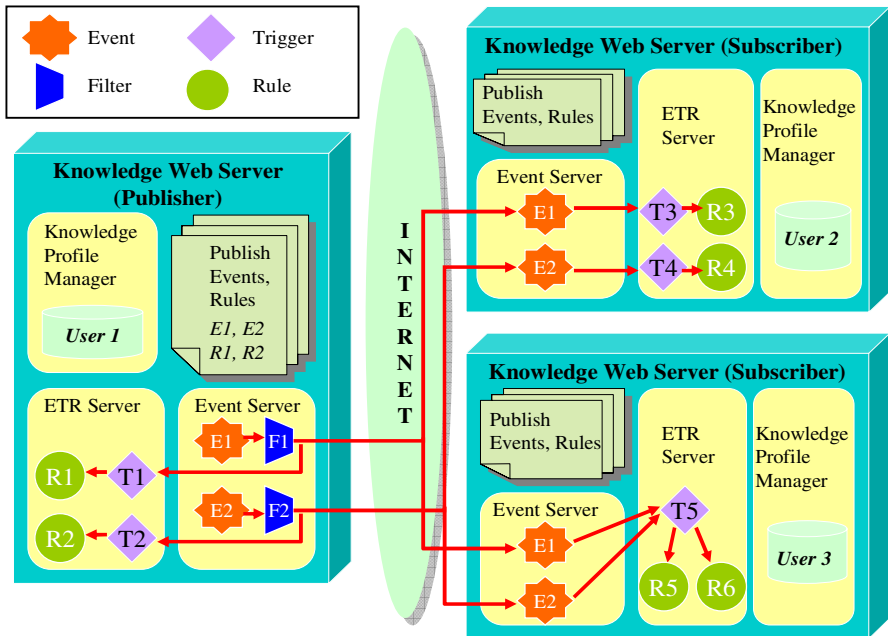


Fig. 9. Architecture of a Knowledge Network

The customizable rule variable allows people to customize rules that are provided by remote Web servers. Event histories and rule structures within the triggers also allow the complex modeling of such relationships that can occur on the Internet. A Knowledge Profile Manager (KPM), which is a Web-based GUI tool that includes event, trigger, and rule editing facilities, was developed to input these knowledge elements into the ETR Server.

Figure 9 shows the architecture of a knowledge network. In such a knowledge network, several knowledge Web servers (KWS's) are interconnected through the Internet. Each KWS contains an Event Server, an ETR Server, and a Knowledge Profile Manager.

In the scenario illustrated by Figure 9, user1 publishes knowledge on a KWS and user2 and user3 are subscribers of the knowledge on other KWS's. Publishers of knowledge may publish events and associated rules on their Web pages. Event filters, which enable subscribers to specify conditions on events and selectively receive event notifications, are also provided. The KPM provides a GUI to perform these operations. User1 has published two events E1 and E2, and two rules R1 and R2. Subscribers of knowledge may visit the publisher's Web page and undergo a registration process to subscribe to the events and also specify filters on the events, and also select published rules to be tied to the event. User2 has subscribed to event E1 and E2. Also, E1 has been linked to rule R1, and E2 has been linked to R2. On the KWS of user2, E1 is linked to rule R3, and E2 is linked to rule R4. User3 has also subscribed to event E1 and E2, and these events are both linked to a complex trigger T5 and again linked to rules R5 and R6 through the trigger.

After this setup stage, events generated on the publisher's site are filtered and notified to subscribers of the events. The rules related to this event are then executed on both the publisher's site and subscriber's site by the ETR Server.

6 Conclusion

Our ETR model supports modeling of inter-organizational systems by providing strong modeling capabilities that differ from conventional distributed system modeling approaches. First, the event specification can effectively support loosely coupled distributed organizations by supporting various types of events including both synchronous and asynchronous events and also allow the delivery of data by providing an arbitrary number of different types of parameters. Second, rules can represent business logic or automatic actions and contain several types of rule variables such as customizable rule variables and existing rule variables to support processing in a distributed environment. Third, in our ETR model, events and rules can be separately defined, modified and added and triggers provide a flexible way of linking events to event history and structures of rules.

The ETR Server has been implemented and can be used for inter-organizational systems such as a Web-based Negotiation Server, a supply chain scenario, a dynamic workflow management system, a Knowledge Network infrastructure, and a transnational information system.

A future research issue would be to provide a mechanism to validate rules that are distributed among organizations in order to avoid situations such as cyclic rule executions or security breaches.

Acknowledgements

The research was supported by the National Science Foundation, USA (Grant #EIA-0075284).

References

1. Su, Stanley Y. W., Huang, C., Hammer, J., Huang, Y., Li, H., Wang, L., Liu, Y., Pluempitiwiriyawej, C., Lee, M., and Lam, H., "An Internet-based Negotiation Server for E-Commerce," *VLDB Journal*, Vol. 10, No. 1, Aug. 2001, pp. 72-90.
2. Su, Stanley Y. W., Lam, H., Lodha, R., Bai, S., and Shen, Z. J., "Collaboration Technologies for Supporting E-supply Chain Management," Chapter in *Applications of Supply Chain Management and E-Commerce Research in Industry*, coedited by E. Akcaly, J. Geunes, P.M. Pardalos, H.E. Romeijn, and Z.J. Shen, Kluwer, 2004.
3. Su, Stanley Y. W., Meng, J., Krishivasan, R., Degwekar, S. and Helal, S., "Dynamic Inter-Enterprise Workflow Management in a Constraint-based E-service Infrastructure," *Electronic Commerce Research*, 3:9-24, 2003.
4. Lee, M., Su, S. Y. W., and Lam, H., "A Web-based Knowledge Network for Supporting Emerging Internet Applications," *WWW Journal*, Vol. 4, No. 1/2, 2001, pp. 121-140.
5. Su, S. Y. W. et al., "A Prototype System for Transnational Information Sharing and Process Coordination," *Proceedings of the National Conference on Digital Government Research (dg.o2004)*, Seattle, Washington, May 24-26, 2004, pp. 199-208.
6. U. Dayal, B.T. Blaustein, A.P. Buchmann, et al. The HiPAC Project: Combining Active Databases and Timing Constraints. In *ACM SIGMOD Record*, Vol. 17(1), March 1988, pp. 51-70.
7. I. Ben-Shaul and S. Ifergan. WebRule: An Event-based Framework for Active Collaboration among Web Servers. In *Computer Networks and ISDN Systems*, Vol. 29 (8-13), October 1997, pp. 1029-1040.
8. A. Carzaniga, D.S. Rosenblum, and A.L. Wolf. Achieving Expressiveness and Scalability in an Internet-Scale Event Notification Service. In *Proc. of the 19th ACM Symposium on Principles of Distributed Computing (PODC2000)*, Portland, OR, July 2000, pp. 219-227.
9. S. Brandt and A. Kristensen. Web Push as an Internet Notification Service. *W3C Workshop on Push Technology*. <http://keryxsoft.hpl.hp.com/doc/ins.html>, Boston, MA, September 1997.
10. Object Management Group (OMG), *CORBA Notification Service*, specification version 1.0. June 20, 2000.
11. Cugola G., Di Nitto E., Fuggetta A., "The JEDI event-based infrastructure and its application to the development of the OPSS WFMS," *IEEE Transactions on Software Engineering*, September 2001, Vol. 27(9), pp. 827 – 850.
12. Robinson R., Rakotonirainy A., "Multimedia customization using an event notification protocol," *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops*, July 2002, pp. 549 – 554.

Semantic LBS: Ontological Approach for Enhancing Interoperability in Location Based Services

Jong-Woo Kim¹, Ju-Yeon Kim¹, and Chang-Soo Kim^{2,*}

¹ Interdisciplinary Program of Information Security,
Pukyong National University, Korea
{jwkim73, jykim}@pknu.ac.kr

² Dept. of Computer Science, Pukyong National University, Korea
cskim@pknu.ac.kr

Abstract. Location Based services (LBS) is a recent concept that integrates a mobile device's location with other information in order to provide added value to a user. Although Location Based Services provide users with much comfortable information, there are some complex issues. One of the most important issue is managing and sharing heterogeneous and numerous data in decentralized environments. The problem makes interoperability among LBS middleware, LBS contents providers, and LBS applications difficult. In this paper, we propose Semantic LBS Model as one of the solution to resolve the problem. Semantic LBS Model is a LBS middleware model that includes a data model for LBS POI information and its processing mechanism based on Semantic Web technologies. Semantic LBS Model provide rich expressiveness, interoperability, flexibility, and advanced POI retrieval services by associating POI Description Language (POIDL) ontology with heterogeneous domain specific ontologies.

1 Introduction

Recently, a lot of researches on the services for supporting the ubiquitous computing are undergoing in the various areas with the growing interest on ubiquitous computing. Context-Sensitive Computing is one of the key technologies supporting the ubiquitous computing which requires the information suitable for the user's context, even in the restricted environment like mobile devices. Location Based Services especially provide the context sensitive information based on the user's location. Location Based services (LBS) is a recent concept that integrates a mobile device's location with other information in order to provide added value to a user [11,13].

Although Location Based Services provide users with much comfortable information, there are some complex issues. One of the most important issue is

* Corresponding author.

managing and sharing heterogeneous and numerous data in decentralized environments. To resolve the problem, several efforts and studies have been made on improving efficiency of data management and establishing standards for information sharing in various operating environments each other [2,3,7].

In this paper we would like to propose Semantic Web approach to enhance interoperability through sharing LBS information. Semantic Web [1] is a technology to add well-defined meaning to information on the Web to enable computer as well as people to understand meaning of the documents easily. For our approach, we propose a Semantic LBS Model that is a LBS middleware model that includes a data model for LBS POI information and its processing mechanism based on Semantic Web technologies. We especially specify POI Description Language (POIDL) ontology that is a ontology-based description language. It can provide interoperability among LBS middleware, LBS contents providers, and LBS applications by allowing POI providers to describe their contents over domain specific ontologies.

In section2, we introduce related work on LBS middleware for interoperability. In section 3, overview of Semantic LBS Model is described. In section 4, the description of LBS ontologies is given. In Section 5, we describe Semantic LBS middleware. Discussion and conclusion is given in Section 6.

2 Related Work

Location Based Services require integration of various technologies and standards. In order to make Location Based Services work, the industry had to overcome several challenges of both a technological and economic nature over the past years. Technologically, realizing LBS can be described by a three-tier communication model [13], including a positioning layer, a middleware layer, and an application layer.

A middleware layer can significantly reduce the complexity of service integration because it is connected to the network and an operator's service environment once and then mitigates and controls all location services added in the future. Moreover, a middleware layer can help LBS applications provide added value related to user's location from heterogeneous data. As a result, it saves operators and third-party application providers time and cost for integrating application. Although most of the commercial LBS platforms have been implemented based on DBMS-based middleware, some approaches to efficiently manage heterogeneous data have been researched [2,3,7].

Although there have been a lot of studies on LBS middleware, LBS middleware has several challenges. First, providing users with added value to mere location information is a complex task, and the basic requirements of the variety LBS applications are numerous [11]. Second, it is difficult to manage LBS contents as a general data management in order to provide users with dynamic information frequently changed [12]. Next, it is difficult to share LBS information because the location-based services are operated in different processing methods, appropriative data exchange protocol, and various platforms [8].

3 Semantic LBS Model

Semantic LBS is a platform that provides novel location-based services that provide not only LBS core services but also more enhanced POI(Point of Interest) retrieval service. A POI is a place, product, or service with a fixed position, typically identified by name rather than by address and characterized by type. A distinguishing feature of Semantic LBS is to retrieve the POIs with domain specific information by providing automatical interaction mechanism based on ontologies.

Figure 1 shows the conceptual service model of Semantic LBS organized for the following three novel services. First, Semantic LBS provide the enhanced LBS Directory Service [10] that retrieves not only the POIs based on user’s location but also their domain specific information. It is difficult for current LBS to retrieve domain specific information because current LBS support searching only information stored as predefined data model. However, because Semantic LBS includes ontology-based data model that allow domain specific information to be specified, it is possible to retrieve domain specific information based on user’s complex requirements. Second, Semantic LBS is able to retrieve real-time updated information. In the Semantic LBS, contents providers publish the contents for each POI, and each POI is stored in decentralized system. The Semantic LBS provides the mechanism that enable agent to retrieve the contents stored in decentralized system and update information in POI repositories for LBS Directory Service [10].

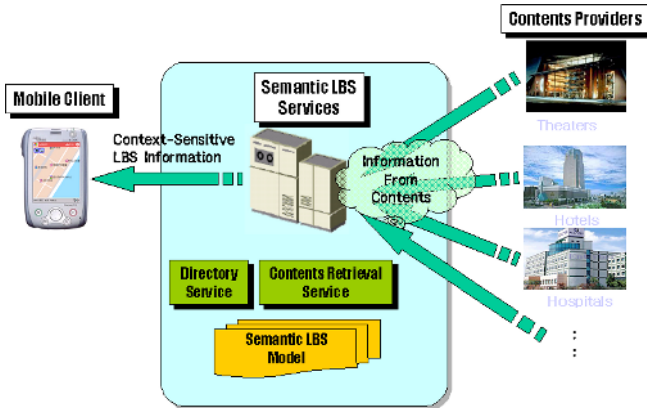


Fig. 1. The conceptual service model of the Semantic LBS

The novel services of Semantic LBS are provided based on Semantic LBS Model. Semantic LBS Model is a LBS middleware model that includes a data model for LBS POI information and its processing mechanism based on Semantic Web technologies. For the data model, we specify two fundamental LBS ontologies for describing POI information: POIDL ontology and LBSTaxonomy ontology. POIDL is a OWL-based description language that allow POI providers to

describe their contents over domain specific ontologies. LBSTaxonomy is an ontology constructed as hierarchical taxonomy of POI types, and it is used to index instances described by the POIDL as POI types. POIs are actually described by the LBS ontologies, various domain specific ontologies, and association between them. Based on the data model, Semantic LBS Model provide some queries to retrieve and update POIs and their heterogeneous domain specific information.

4 Modeling Ontologies for LBS POIs

This section presents ontologies for describing LBS POIs (Points of Interest), and details some of the key classes and properties. LBS ontologies is composed of three kinds of ontologies: POIDL ontology, LBSTaxonomy ontology, and domain ontologies each with their unique XML namespace.

POIDL ontology is most fundamental ontology in order to specify POIs, and it provides sufficient expressiveness which can specify not only general information provided in current LBS model but also domain specific information. LBSTaxonomy ontology is an ontology which hierarchically classifies LBS POIs by types of services. It includes only necessary information for retrieving POIs based on classification and location. Domain ontologies are ontologies to specify domain specific information, and are used so that each contents provider specifies pertinent POI information conforming with its characteristics.

4.1 POIDL Ontology

POIDL is an ontology-based POI(Point of Interest) specification language that allows LBS contents providers to describe their information over domain specific ontologies. In addition, POIDL allow the information for query used to search and update domain specific information to be specified. This mechanism enable LBS to provide POI information changed frequently. POIDL ontology includes three classes: POI class, Content class, and ContentQuery class.

POI class is used to describe general information about a POI and its domain-dependent information. Figure 2 is an example of describing a POI using POI class. Instances of POI class specify POI information using five properties as follows.

- `poidl:name`. This specifies a name of the POI that is specified in the instance of the POI class to identify the POI. In Figure 2, this property specifies that the instance describe the POI that has name of 'HotelA'.
- `poidl:isIncludeIn`. : This property specifies a type of the POI instance. This associates the POI instance with a class of LBSTaxonomy ontology described in Section 4.2. The LBSTaxonomy ontology is an ontology that specifies hierarchical classification of POI types. In Figure 2, this property specifies that the type of the POI instance is 'Hotel'.
- `poidl:hasContent`. This property is used to associate an instance of the Content class that specifies domain specific information about the POI, and a POI can have more than one contents. In Figure 2, this property associates

```

<poidl:POI rdf:ID="POI_HotelA">
  <poidl:name>HotelA</poidl:name>
  <poidl:isIncludedIn rdf:resource="#&#x2013;staxonomy;Hotel">
  <poidl:hasContent rdf:resource="#Grade_HotelA">
  <poidl:hasContent rdf:resource="#EmptyRoom_HotelA">
  <poidl:rectLeft>210415</poidl:rectLeft>
  <poidl:rectRight>210415</poidl:rectRight>
  <poidl:rectTop>1270223</poidl:rectTop>
  <poidl:rectBottom>127323</poidl:rectBottom>
</poidl:POI>

```

Fig. 2. An example instance of POI class

two contents, 'Grade_HotelA' and "EmptRoom_HotelA", that are instances of Content class.

- The four location properties - poidl:rectLeft, poidl:rectRight, poidli:rectTop, and poidl:rectBottom describe the boundary of a POI. Location can be described by various form, but we suppose that location is represented by rectangular coordinates used in geographical map. In order to indicate geographical location through various coordinates systems and address, we can specify an ontology for location and associate an instance of location ontology.

Content class is used to describe a domain specific information for a POI, and an instance of the Content class specifies type and value of the POI. The Content class provides rich expressiveness and flexibility because instances of the class specify domain specific information through referring domain specific ontologies and its instances. Also the Content class associates an instance of the ContentQuery class that specifies a query for retrieving current value in its domain. Figure 3 is two example instances of the Content class, which describe grade of 'HotelA' and the number of empty room of 'HotelA' respectively. the grade is static information, while the number of empty room is a dynamic information that changes frequently. The three properties for describing an instance of the Content class is as follows.

- poidl:typeofcontent. This property specifies a type of domain specific information and associates a class of the domain specific ontology of the POI. In Figure 3, this property describe that an instance of the Content class, 'Grade_HotelA' specifies the grade of 'HotelA'.
- poidl:valueofcontent. This specifies a value of domain specific information of the POI. In example of Figure 3, This property specifies that the grade of 'HotelA' is 'Grade 1'. And we can know that the number of empty rooms has to be retrieved because it is dynamic information that changes frequently.
- poidl:query. This associates an instance of ContentQuery class for querying domain specific information of the POI. For example, in order to retrieve

```

<poidl:Content rdf:ID="Grade_HotelA">
  <poidl:typeofcontent rdf:resource="&onhotel;grade">
  <poidl:valueofcontent>Grade 1</poidl:valueofcontent>
  <poidl:query rdf:resource="#Query_Grade_HotelA">
</poidl:POI>

<poidl:Content rdf:ID="EmptyRoom_HotelA">
  <poidl:query rdf:resource="#Query_EmptyRoom__HotelA">
</poidl:POI>

```

Fig. 3. An example instance of Content class

```

<poidl:ContentQuery rdf:ID="Query_EmptyRoom__HotelA">
  <poidl:subject>hotela:HotelA</ibspoi:subject>
  <poidl:predicate>hotela:emptyRoom</ibspoi:predicate>
  <poidl:object>?num</ibspoi:object>
</ibspoi:POI>

```

Fig. 4. An example instance of ContentQuery class

the number of empty rooms for 'HotelA', an instance of ContentQuery class, 'Query_EmptyRoom_HotelA', can be referred.

The ContentQuery class is used to describe a query for retrieving domain specific information from its domain. This class enable domain specific information to be retrieved and automatically updated. The query language to retrieve information from RDF documents such as RDQL is represented as triple, same as RDF. The ContentQuery class specifies a triple for a query - subject, predicate, and object. Figure 4 is an example instance of the ContentQuery class. The three properties for describing an instance of the ContentQuery class is as follows.

- poidl:subject. This property describes a subject of triple for query.
- poidl:predicate. This describes a predicate of triple for query.
- poidl:object. This property describes an object of triple for query.

4.2 LBSTaxonomy Ontology

The LBSTaxonomy ontology is used to guide the taxonomy of POI entities in the LBS domain. An instance of the LBSTaxonomy ontology specifies the information about the POI used in Directory Service of Semantic LBS. An instance of the LBSTaxonomy ontology includes the basic information of the POI such as

```

<tax:Hotel rdf:ID="HotelA">
  <tax:name>HotelA</poidl:name>
  <tax:details rdf:resource="&poidl:POI_HotelA">
  <tax:rectLeft>210415</tax:rectLeft>
  <tax:rectRight>210415</tax:rectRight>
  <tax:rectTop>1270323</tax:rectTop>
  <tax:rectBottom>127323</tax:rectBottom>
</tax:POI>

```

Fig. 5. An Example instance of LBSTaxonomy ontology

name, and location information to efficiently retrieve POIs as taxonomy. It also specifies the information to refer the instance of the POIDL ontology that specifies domain specific information of the POI and association information used to specify general retrieval patterns of user. In this work, the LBSTaxonomy ontology refers to the hierarchical classification code defined by National Geographic Information Institute (NGI) in Republic of Korea, and we extend it. For example, Hotel class is a subclass of ServiceFacility class, and ServiceFacility is a subclass of Structure. Figure 5 shows an example instance of the LBSTaxonomy. LBSTaxonomy ontology includes seven properties as follows.

- tax:details. This property associates an instance of the POI class in POIDL ontology. This property provide connectivity with detailed information for POIs.
- tax:name. This describes identification of a POI. The object of this property is same as the object of 'name' property of POIDL ontology. For example, the 'HotelA' in Figure 5 indicates the identification of a POI.
- The four location properties - tax:rectLeft, tax:rectRight, tax:rectTop, and tax:rectBottom describe the boundary of a POI. The object of this property is same as the object of location properties of POIDL ontology.

LBSTaxonomy ontology is specified to be basically used in Semantic LBS Model, but applications can specify application-specific ontologies as their purposes and features.

4.3 Domain Specific Ontology

The domain specific ontologies define concepts for each domain and relationship between them. The domain specific ontologies can be defined by contents provides or the Standard Organizations. Because LBS contents is different as their domain, the domain specific ontologies help domain specific information to be represented more exactly. The domain specific ontologies is referred by the instances of POIDL ontology to specify the domain specific information.

5 Semantic LBS Directory Service with LBS Ontologies

We approach the enhancement of LBS Directory Service by building the middleware that provides retrieval services for Semantic LBS data model as shown in Figure 6. Semantic LBS Model provides not only general POI retrieval queries but also some advanced queries based on expressiveness and interoperability of Semantic LBS data model. Semantic LBS Model enhances general queries of the Directory Service. The Directory Service of location-based services provides a search capacity for one or more Points of Interest (POI), and it provides several kinds of queries as range of retrieving POIs. Semantic LBS Model enable to retrieve POIs and their domain specific information with more complex conditions, while current LBS Models provide a simple search capacity that can retrieve POIs based on only location. The Semantic LBS Model also provides a contents query that retrieves more detailed and domain specific information about a POI. The query enable to acquire dynamic information changed frequently and automatically update POI directories.

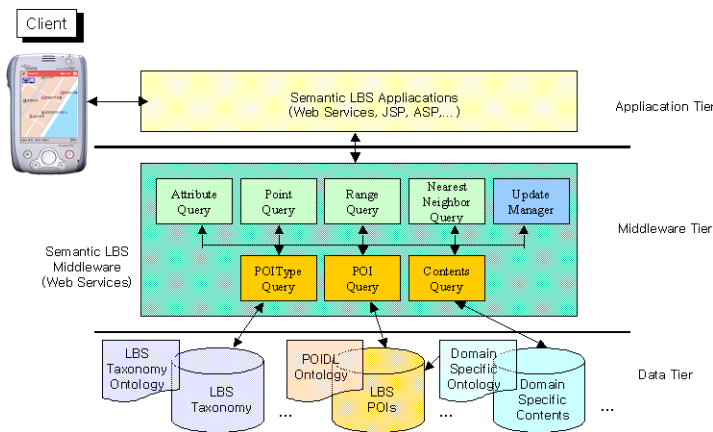


Fig. 6. The architecture of Semantic LBS Middleware

The query modules in Figure 6 are implemented using Jena, Joseki and RDQL templates. Jena API is a Java application programming interface that creates and manipulates RDF documents, and Joseki is a Java client and a server that implements the Jena network API over HTTP [4,9]. We can semantically search the instances of RDF documents through RDQL, a Query Language for RDF, which is Jena’s query language [4].

Figure 7 shows a simple application of Semantic LBS, Hotel Finder. The application provides hotel information including domain specific information such as room type, meal type, and price (Figure 7 (b)). It also provides the map service that utilize Mobile GIS module developed in our previous work (Figure 7 (c)) [5,6].

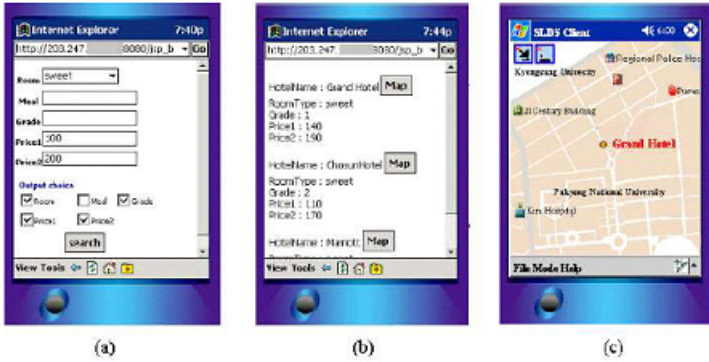


Fig. 7. An example application of Semantic LBS

6 Discussion and Conclusion

Mobile users expect that Location Based Services provide more specific information exactly. Therefore, LBS middleware models have to provide domain specific information even if LBS data is heterogeneous and numerous. Semantic LBS Model is a LBS middleware model that includes a data model for LBS POI information and its processing mechanism based on Semantic Web technologies. For the data model, we specified two fundamental LBS ontologies for describing POI information: POIDL ontology and LBSTaxonomy ontology. POIDL is a OWL-based description language that allow POI providers to describe their contents over domain specific ontologies. LBSTaxonomy is an ontology constructed as hierarchical taxonomy of POI types, and it is used to index instances described by the POIDL as POI types. POIs are actually described by the LBS ontologies and heterogeneous domain specific ontologies. Based on the data model, Semantic LBS Model provide some queries to retrieve and update POIs and their domain specific information. Main contributions of our approach include:

- Expressiveness of POIs: Semantic LBS Model provides sufficient expressiveness. The POIDL in Semantic LBS Model provide more expressive and more flexible description mechanism for POI information that enable domain specific information to be described.
- Interoperability: Semantic LBS model supports interoperability through information sharing in decentralized environments. Semantic LBS Model uses HTTP that is standard data exchange protocol of Web and shares LBS information through URI and ontologies.
- Benefits of POI retrieval: LBS Model provides not only basic queries for retrieving POIs but also some advanced mechanism that retrieves POI information: retrieving domain specific information even if the information changes frequently, and automatically updating domain specific information of POIs for LBS directory service. The ontology-based data model enable Semantic LBS to provide the advanced functions.

- Flexibility: Semantic LBS allow the domain specific ontologies to extend without modifying the middleware and applications. Moreover, because POIDL specifies templates for retrieving domain specific information of each POI, Semantic LBS is able to append POI information easily.

Acknowledgement. This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD) (KRF-2005-908-D00057).

References

1. Berners-Lee, T., and et. al.: The Semantic Web, Scientific American (2001)
2. I. Burcea and H.-A. Jacobsen: L-ToPSS: Push-Oriented Location-Based Services, In Proc. of the 4th VLDB Workshop on Technologies for E-Services (TES'03), Lecture Notes in Computer Science, No. 2819, Springer-Verlag (2003)
3. N. Davies, S.P. Wade, A. Friday, and G. S. Blair: Limbo: A Tuple Space Based Platform for Adaptive Mobile Applications, In Proceedings of the International Conference on Open Distributed Processing/Distributed Platforms (ICODP/ICDP) (1997)
4. Jeremy J. and et. al.: Jena: Implementing the Semantic Web Recommendation, ACM (2004)
5. J.W. Kim, et. al.: The Efficient Web-Based Mobile GIS Service System through Reduction of Digital Map. Lecture Notes in Computer Science, Vol. 3043, Springer-Verlag (2004)
6. J.W. Kim, et. al.: Location-Based Tour Guide System Using Mobile GIS and Web Crawling, Lecture Notes in Computer Science, Vol. 3428, Springer-Verlag (2005)
7. H. Leung, I. Burcea, and H.-A. Jacobsen; Modeling Location-based Services With Subject Spaces, In Proceedings of CASCON Conference, IBM Publisher (2003)
8. Mabrouk, M. et. al.: OpenGIS Location Services (OpenLS): Core Services, OpenGIS Implementation Specification, OGC (2004)
9. McBride, B.: Jena: A Semantic Web Toolkit. IEEE INTERNET COMPUTING. IEEE (2002)
10. OpenLS Initiative.: A Request for Technology in Support of an Open Location Services Testbed. OGC (2000)
11. J. Schiller: Location-Based Services, Elsevier, p.1-5 (2005)
12. Shioyang Wu, Kun-Ta Wu.: Dynamic Data Management for Location Based Services in Mobile Environments. Proc. 2003 Int. Conf. Database Engineering and Applications symposium. IEEE (2003)
13. S. Spiekermann: General Aspects of Location-Based Services, Location-Based Services, Elsevier 2005.

Dynamic Consistency Between Value and Coordination Models – Research Issues

Lianne Bodenstaff*, Andreas Wombacher, and Manfred Reichert

Information Systems Group, Department of Computer Science,
University of Twente, The Netherlands
{l.bodenstaff, a.wombacher, m.u.reichert}@utwente.nl

Abstract. Inter-organizational business cooperations can be described from different viewpoints each fulfilling a specific purpose. Since all viewpoints describe the same system they must not contradict each other, thus, must be consistent. Consistency can be checked based on common semantic concepts of the different viewpoints. This is sufficient for equal concepts, while weakly related concepts, e.g. related to runtime behavior of viewpoints, have to be considered explicitly. In this paper we identify dynamic consistency issues correlated to the runtime behavior between value and coordination viewpoints on behalf of an example. In particular, an issue class on occurrence estimations of execution options and an issue class on granularity differences in modelling are identified and illustrated.

1 Introduction

Modelling inter-organizational business cooperations constitutes a crucial task that can be done from different viewpoints. Each viewpoint emphasizes an important aspect of the cooperation. In this paper two viewpoints are dealt with.

The *value viewpoint* gives an indication on the profitability of the cooperation. The value model describing this viewpoint models which objects of economic value are exchanged between parties. Furthermore, estimations, e.g. on the number of occurrences of an object of value, are modelled. The value model enables talking about the commercial interests of the different business actors and abstracts from processes and object flow, i.e., it models *what* objects of value are exchanged but not *how* this exchange is realized. The *coordination viewpoint*, in turn, represents the interactions and interdependencies between the cooperating parties in terms of exchanged messages. The model describing the coordination viewpoint represents *how* the actors in the model cooperate, i.e., it represents the coordination of the exchanges. Together, the two viewpoints describe *what* is exchanged of value between the parties and *how* these exchanges can be realized.

Multi-viewpoint descriptions of complex systems must maintain consistency across viewpoints. To ensure both models indeed describe the same cooperation,

* Supported by the Netherlands Organisation for Scientific Research (NWO) under contract number 612.063.409 (Value-Based IT Alignment).

they have to be checked for consistency, i.e., we have to validate that the overlapping system specification contained in both viewpoints is not contradicting.

Our work will build on the approach presented in [1]. So far, this approach has solely considered consistency checking of static aspects, i.e., during design time, and does not consider the runtime behavior of a model. Therefore, certain aspects of a model, e.g. estimations made in the value model, are not considered. However, these estimations should still be consistent with the dynamic aspects of the coordination model. In this paper we refer to consistency of the static aspects as *static consistency* and consistency of the dynamic aspects will be referred to as *dynamic consistency*.

To illustrate relevant issues, we use a running example in which we abstract from details for the sake of simplicity. This example consists of a health insurance company which provides one-year insurance to its customers based on monthly paid premiums. Insured customers can claim refunds for treatments they paid themselves. Furthermore, the insurance company gets money from CVZ for every paid refund to the customer. CVZ is a Dutch organization distributing tax money from the government to the insurance companies. CVZ gets funding on an annual basis in exchange for a proof of proper distribution of tax money.

The paper is structured as follows: Section 2 and Section 3 explain in detail value and coordination modelling. After that consistency aspects are discussed in Section 4. Section 5 identifies research issues in dynamic consistency checking between the value and coordination models. In Section 6 we discuss related work. We end this paper with a summary and outlook in Section 7.

2 Value Model

For inter-organizational design the value viewpoint is especially important because all actors involved are profit-and-loss responsible. The expected revenue for every actor is calculated through a method of cost-benefit analysis (like e.g. Net Present Value (NPV) [2], Return on Investment [3] and Real Options Analysis [4]). In this paper we use e^3 -value [5] because of its graphical representation. However, the issues raised in this paper apply to value models in general. e^3 -value uses NPV for cost-benefit analysis.

We informally describe the semantics of basic e^3 -value concepts [5], based on Figure 1. It depicts our sample business case as explained in the introduction as an e^3 -value model. The example depicts four *actors* and eight *value transfers*. For example, one value object, *premium*, is transferred from the customer to the insurance company. Another value object, the *insurance* itself, is transferred from the insurance company to the customer. These two transfers are in Figure 1 annotated with an ‘F’. A combination of value transfers in one transaction is referred to as a *value exchange*. In e^3 -value a distinction is made between different kinds of value objects. A value object is either a *product*, *service*, *money* or *consumer experience*. In this example the *premium* is a value object of the *money* type and the *insurance* provided by the insurance company can be considered as a *service*.

The consumer need is “having a health insurance for one year”. This is represented by placing the *start stimulus* at the customer. Now, the set of value objects that needs to be transferred to fulfill the consumer need, consists of all value transfers connected through the *dependency path* in the model. Every month there are two possible sets of value transfers that can fulfill the consumer need. Either the customer claims restitution for *treatments* he paid for himself and he pays the monthly *premium*, or he only pays the monthly *premium*. When the customer claims a restitution, the insurance company claims compensation from CVZ. CVZ, in turn, gets its *funding* from the government. The health insurance company has multiple customers, represented as a *market segment* in the figure.

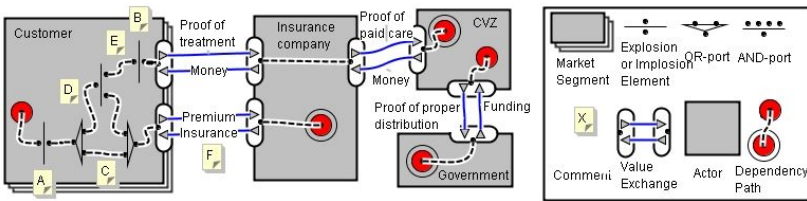


Fig. 1. e³-value model, business case

In Figure 1, the twelve monthly payments for fulfilling one consumer need are realized by adding an *explosion element*, annotated with ‘A’ in the figure, associated with ratio 1 : 12. The choice between the two options for fulfilling the consumer need is represented as an OR-split in the figure. After an OR-split only one of the dependency paths is chosen. When the customer has not received treatments that month, the path annotated with ‘C’ is chosen. The two resulting value transfers constitute the first set of transfers that can fulfill the consumer need. If the customer did receive treatment that month, the path annotated with ‘D’ is chosen. This path splits through an AND-split, representing a parallel occurrence of two or more dependency paths. In an AND-join all entering dependency paths share the continuation of the dependency path. Now, both value exchanges in the model between the insurance company and the customer occur. To enable more than one restitution claim per month another explosion element, annotated with ‘B’, is added. The insurance company claims restitution from CVZ. These value transfers together, are the second set of value transfers. The dependency path starting within CVZ represents the third set of value transfers. In Section 5.4 the reason for intermitting the dependency path is explained.

In the profitability sheets, associated with the graphical representation of the value model, the estimations are denoted. The market segment is quantified by estimating the number of customers and the ratio on the explosion elements and OR-split is set. For every monetary value transfer the quantification is denoted in the profitability sheets. Now, the expected revenue for every actor in the model can be calculated.

3 Coordination Model

In a cooperation the messages between actors are exchanged in a particular order which is not represented in the value model. The set of ordered tasks and message exchanges is referred to as the *execution sequence*. The coordination model is important to determine conceptual problems of the cooperation at an early stage. Coordination model examples are e.g. Finite State Automata (FSA), Petri Nets [6], Workflow Nets and flowcharts.

In this paper we use Petri Nets [7] to represent coordination models because of its graphical representation, formal semantics and the variety of available tools. Although we use Petri Nets in this paper, the issues illustrated are modelling technique independent. In Figure 2 the sample business case as described in the introduction is represented as a coordination model in terms of a Petri Net. First, the basic concepts of a Petri Net are introduced after which the business case is explained in more detail.

The static part of the Petri Net consists of *places* (indicated as circles) and *transitions* (indicated as rectangles) which are connected with each other through *arcs*. Places represent message exchanges and transitions represent tasks. Furthermore, the dynamic part of the Petri Net enables simulation of executions in the model. A place can hold zero or more tokens. A distribution of tokens over places represents the state of a Petri Net. A transition is called enabled, i.e., it may fire, if each place connected to the transition with an incoming arc, holds at least one token. When a transition fires a token is removed from each of these places and a token is put in every place connected with the transition

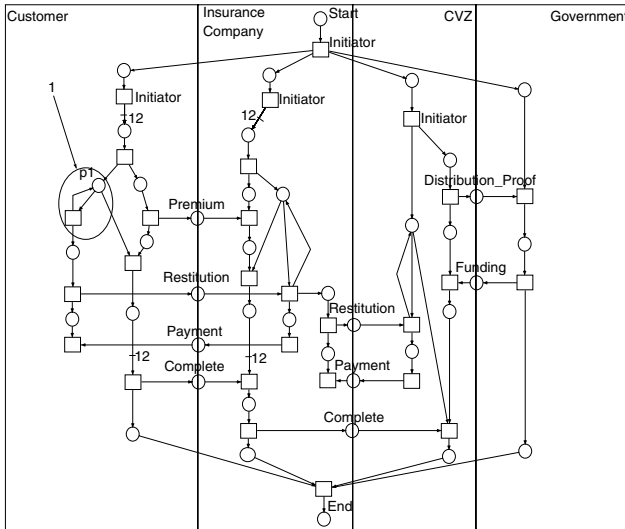


Fig. 2. Coordination model in terms of a Petri Net

through an outgoing arc. A Petri Net and a distribution of tokens over places is also referred to as an *instance of a coordination model*.

The message exchanges are modelled as places on the border between two actors. After a token is put at the start place the first *initiator* transition is enabled. If the transition fires, it enables *initiator* transitions for every actor. The customer has two parallel execution sequences. The first ensures the message exchange of the payment of the *premium* to the insurance company. The second sequence depicts the message exchanges of asking for a *restitution* to the insurance company and of receiving the restitution *payment* from the insurance company. A recursive process, annotated with ‘1’ in Figure 2, is used to allow claiming more than one restitution per month. After the customer has paid all monthly premiums and received all restitution payments from the insurance company, he sends a *complete* message to the insurance company. For every request of *restitution* by the customer, the insurance company sends a *payment* to the customer as well as a request for *restitution* to CVZ. After the insurance company received all payments from CVZ and has performed all payments to the customer, it sends a *complete* message to CVZ. CVZ receives, after sending a message with proof of proper distribution, *funding* for a year by the government. In parallel, CVZ receives messages from the insurance company for *restitutions* and pays the restitutions to the insurance company. After the insurance company exchanges the *complete* message and CVZ has received *funding* from the government, a token is available in every place needed for enabling the *end* transition to terminate the process.

4 Consistency Between Value and Coordination Models

The value model and coordination model sketched in the previous sections describe the same system from different viewpoints. To ensure that both models indeed are related to the same system, we have to check whether these two viewpoints are consistent with each other. In [1] an intuitive definition of consistency between a value and a coordination model has been defined.

A value and coordination model are considered to be consistent if:

1. for every set of value transfers in the value model (dependency path in e^3 -value), there exists an execution sequence in the coordination model such that exactly the product/money value transfers contained in the set are exchanged in the execution sequence, and
2. for every execution sequence in the coordination model, there exists a set of value transfers in the value model (dependency path in e^3 -value) such that the message exchanges contained in the execution sequence represent product/money value transfers exchanged in the set of value transfers.

Note that the definition focusses on product and money value transfers, since experience and service value transfers are not instantiating an explicit message exchange. A message exchange represents a product value transfer, if the sender and receiver of the message exchange equals the provider and the recipient of

a value. With regard to the example in Section 2 the *insurance* value transfer is a service, thus, can not be correlated with a message exchange. The money value transfer *premium* in the value model is provided by the *customer* and received by the *insurance company*. A corresponding message exchange sent by the *customer* and received by the *insurance company* is also contained in the coordination model.

The consistency definition mentioned so far ignores the dynamics of the modelled system, resulting in estimations in the value model and observed behavior in the coordination model.

5 Research Issues

In this section we demonstrate the need for dynamic consistency checking by identifying major consistency issues that occur during runtime and could not be identified during design time. We identify two classes of issues that concern mismatches between value and coordination model.

The first class concerns a mismatch between the **estimations** made in the profitability analysis and the execution semantics of the coordination model. This class represents the mismatch between the estimated number of occurrences and choices between sets of transfers in the value model and actual occurrences and choices of message exchanges in the coordination model. The second class deals with the mismatch of different levels of **granularity**. Model boundaries can vary among models with different purposes although describing the same system. Furthermore, within a model different levels of granularity can occur. This class covers mismatches of granularity differences between actors and value transfers in the value model itself as well as between the value and coordination model. Next, the issues are illustrated by the use of our example.

5.1 Issue 1: Number of Occurrences of a Value Transfer

For the fulfillment of one consumer need, a specific value transfer might occur several times. This number of occurrences may be fixed or it can be an estimated average of occurrences of value transfers over periods of time and actors. When estimating the profitability of the cooperation in the value model, the number of expected occurrences of each value transfer compared to a single consumer need as well as the value of each transfer is estimated.

As an example, in e^3 -value the ratio between the consumer need and a value transfer as well as the ratio between two value transfers is represented by an explosion or implosion element. Regarding our example, the consumer need will be fulfilled if the *premium* is paid twelve consecutive months. In the value model, Figure 3(a), this is modelled by adding an explosion element with ratio 1 : 12. We denote this ratio as a *fixed ratio* because it is the same with every customer and every case. Furthermore, a customer uses its insurance by asking one or more restitutions. This is again modelled as an explosion element with, in this example, an associated ratio of 1 : 1, 5. We denote this type of ratio as an *average ratio*.

The coordination model must contain a correspondence to the number of occurrences of value transfers as expressed in the value model. In the coordination model a value transfer with a fixed ratio is represented by forcing a fixed number of message exchanges to occur. In the case of an average ratio, a construction for enabling repetitions of tasks is used.

Using a Petri Net-based coordination model, the fixed ratio of monthly payments can be realized by adding an initiator transition for the customer. The initiator transition inserts twelve tokens for further processing of premiums and restitutions, represented in the upper part of Figure 3(b). Furthermore, to allow claiming more than one restitution per month a recursive process, annotated with ‘1’ in Figure 2, is used. In the coordination model there is no ratio represented between the monthly paid premium and the amount of restitutions as it is done in the profitability analysis of the value model, highlighted in the lower part of Figure 3(b).

In case of having fixed ratios consistency can be assured when designing the models while having average ratios is an example of the first class of issues.

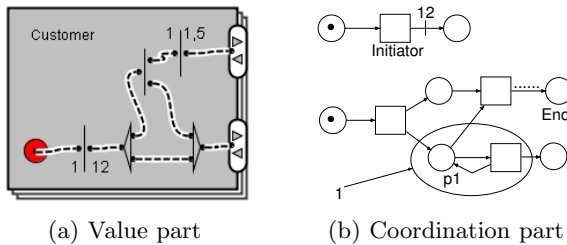
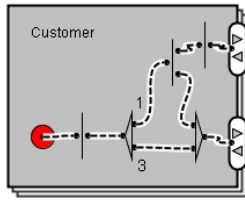


Fig. 3. Illustration of Issue 1

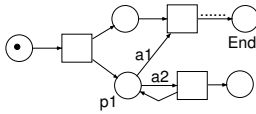
5.2 Issue 2: Choices in Sets of Value Transfers

If a consumer need can be fulfilled in multiple ways by carrying out different sets of value transfers, each of these sets is represented in the value model. Further, every set is associated with an expected percentage of consumer needs that will be fulfilled by using that specific set. In the example (see Figure 1), the consumer need can be fulfilled by either using the possibility of restitution during a monthly period or by not using this possibility. The ratio between these options is estimated by the insurance company based on all its customers and their restitution requests in previous years. This is again modelled in Figure 4(a) where the estimated ratio is modelled as 1 : 3.

In the coordination model the different ways of fulfilling a consumer need are represented as decisions between tasks. In our Petri Net (see Figure 2), for example, the decision point of requesting a restitution is place *p1* (see mark ‘1’). In Figure 4(b), this part of the Petri Net is again depicted where the ratio between arc *a1* and *a2* is not represented. Now the mismatch is that the profitability analysis of the value model is based on an average over a specific period of time while during runtime of the coordination model either restitutions occur or not.

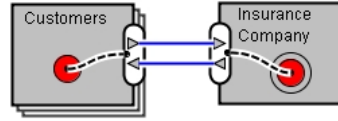


(a) Value part

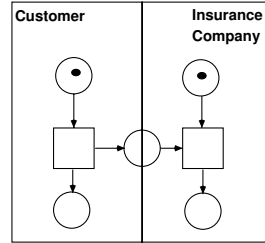


(b) Coordination part

Fig. 4. Illustration of Issue 2



(a) Value part



(b) Coordination part

Fig. 5. Illustration of Issue 3

The average value can only be determined after the coordination process has been executed several times.

Checking estimated ratios on choices in sets of value transfers with runtime instances of the coordination model, belongs to the first class of issues.

5.3 Issue 3: Granularity Difference Between Actors

As another issue, calculation methods in value modelling use estimations on the number of occurrences of value transfers based on groups of actors. However, in the coordination model every actor is modelled separately. Thus, the estimated number of occurrences of value transfers based on an actor group in the value model cannot be directly related to the real time number of occurrences of message exchanges per actor in the coordination model.

In the value model, for example, the insurance company interacts with several customers rather than a single actor. However, the coordination model represents the interaction between the insurance company and a single customer. Thus, the two models have different levels of granularity of actors. This is an issue for dynamic consistency checking because the average of restitutions in the value model can only be compared with the average value calculated over several instances associated to different actors of a coordination model.

A schematic example of this issue is given in Figure 5. The coordination model captures only a *fraction* of the market segment represented in the value model, i.e., one customer.

5.4 Issue 4: Granularity Difference Between Value Transfers

Recall that the purpose of a model determines which information is represented in a model. If, due to the boundaries of the model, one value transfer represents

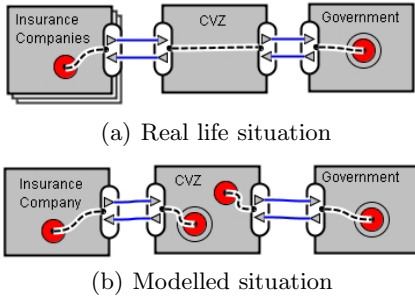


Fig. 6. Value model, Issue 4

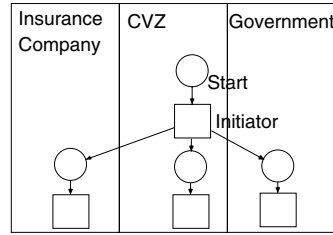


Fig. 7. Coordination model, Issue 4

the transfer of a value object concerning a market segment while another transfer represents the transfer of a value object concerning one specific actor of that market segment, a granularity difference between these value transfers occurs. Since a value transfer is atomic and therefore can not be partly executed, the relation between both value transfers cannot be straightforwardly calculated.

For example, for paying restitutions to all insurance companies, CVZ gets a fixed amount of funding from the government. This is a value transfer concerning the market segment of insurance companies. The purpose of the value model is to estimate the revenue of one specific actor of the market segment insurance companies. Therefore, only a fraction of the value transfer between the government and CVZ is relevant for our modelling purpose. This results in a different level of granularity between the value transfers between CVZ and the government, concerning the market segment, and the value transfers between CVZ and the particular insurance company. Therefore, the two interfaces of CVZ cannot be related through a dependency path and thus the dependency path is broken within actor CVZ as depicted in Figures 6 and 7.

In the coordination model this granularity difference is not present because every actor is modelled separately. Thus, there is a mismatch between the granularity in the value model affecting consistency with the coordination model.

Recall that the definition of static consistency is based upon matching the execution paths in the coordination model and the dependency paths in the value model (cf. Section 4). The separation of the dependency paths creates two independent paths which must be related to a single execution path in the coordination model. In the profitability analysis, however, there is a relation between both dependency paths. Thus, the estimations made in the profitability analysis on the relation of the dependency paths have to be checked after runtime of the coordination model.

6 Related Work

Consistency between different viewpoints is an important issue addressed often in literature. In particular, there exist different ways of defining consistency within a single viewpoint as well as between different viewpoints. For instance in the

workflow community different notions of consistency mainly based on deadlock-freeness have been defined on all kinds of workflow models, like e.g. Workflow Nets, guarded Finite State Automata, Coloured Place/Transition Nets, or statecharts. Further there exist proposals to extend consistency between different models of the same viewpoint again focusing on deadlock-freeness like e.g. [8,9,10] for the different models.

Consistency between different viewpoints has been addressed on different levels of abstraction. An analysis on the conceptual level has been provided in [11] where the value and coordination viewpoints are compared based on the semantic concepts used in the different viewpoints. A human intuitive consistency definition has been proposed in [12] which gives an understanding on what consistency means without explaining how to check it. This intuitive definition has been operationalized in [1]. However, this consistency definition does not consider dynamic consistency.

Besides the above mentioned approaches on checking consistency between viewpoints, there exist constructive approaches guaranteeing consistency of the model derived from another model. For example in [13] an approach is proposed to use an intermediate model as a bridge between a business model and a process model. [14] propose a chaining method to derive from a business model a corresponding process model. The approach is based on associating different value transfer to off-the-shelf process patterns and combining these patterns. All these constructive approaches focus on static consistency and do not address the issues raised in this paper.

7 Summary and Outlook

In this paper we illustrate issues related to dynamic consistency checking between value and coordination models by the use of concrete examples. More specifically, we identify two classes of major consistency issues. Furthermore, we illustrated the need for a dynamic consistency definition, since current consistency definitions, e.g. as defined in [1], cannot check consistency between two models for all aspects, i.e., dynamic as well as static aspects. The contribution of this paper is the identification and structuring of these issues. We continue this research by investigating a dynamic consistency definition to resolve the issues raised in this paper.

The authors thank Roel Wieringa and Jaap Gordijn for participating in discussions and giving their comments on earlier versions of this paper.

References

1. Zlatev, Z., Wombacher, A.: Consistency between e^3 -value models and activity diagrams in a multi-perspective development method. In: OTM Conferences (1). (2005) 520–538
2. Laudon, K., Laudon, J.: Essentials of Management Information Systems. 5 edn. Prentice Hall (2003)

3. Friedlob, G.T., Plewa Jr., F.J.: Understanding Return on Investment. John Wiley & Sons, Inc. (1996)
4. Benaroch, M.: Managing information technology investment risk: A real options perspective. *Journal of Management Information Systems* **19**(2) (2002) 43–84
5. Gordijn, J., Akkermans, J.M.: Value-based requirements engineering: Exploring innovative e-commerce ideas. *Requirements Engineering* **8**(2) (2003) 114–134
6. Peterson, J.L.: Petri Net Theory and the Modeling of Systems. Prentice-Hall (1981)
7. Jensen, K.: Coloured Petri Nets. Basic Concepts, Analysis Methods and Practical Use. Springer (1997) Three Volumes.
8. Aalst, W., Weske, M.: The P2P approach to interorganizational workflows. In: Proceedings of 13. International Conference on Advanced Information Systems Engineering (CAISE), Interlaken, Switzerland (2001)
9. Wombacher, A., Fankhauser, P., Aberer, K.: Overview on decentralized establishment of consistent multi-lateral collaborations based on asynchronous communication. In: Proc. IEEE Int'l. Conf. on e-Technology, e-Commerce and e-Service (EEE). (2005) 164–170
10. Kindler, E., Martens, A., Reisig, W.: Interoperability of workflow applications: Local criteria for global soundness. In: Business Process Management, Models, Techniques, and Empirical Studies. (2000) 235–253
11. Gordijn, J., Akkermans, J., van Vliet, J.: Business modelling is not process modelling. In: Conceptual Modeling for E-Business and the Web. Volume 1921., Springer LNCS (2000) 40–51
12. Wieringa, R.J., Gordijn, J.: Value-oriented design of service coordination processes: correctness and trust. In: Proceedings of the ACM Symposium on Applied Computing (SAC, New York, NY, USA, ACM Press (2005) 1320–1327
13. Andersson, B., Bergholtz, M., Edirisuriya, A., Ilayperuma, T., Johannesson, P.: A declarative foundation of process models. In: Proc. of the 17th International Conference on Advanced Information Systems Engineering. (2005) 233–247
14. Andersson, B., Bergholtz, M., Grégoire, B., Johannesson, P., Schmitt, M., Zdravkovic, J.: From business to process models - a chaining methodology. In: CAiSE2006: Proceedings of the 18th International Conference on Advanced Information Systems Engineering, Luxembourg (2006) 211–218

PROMONT – A Project Management Ontology as a Reference for Virtual Project Organizations

Sven Abels³, Frederik Ahlemann², Axel Hahn¹, Kevin Hausmann¹,
and Jan Strickmann⁴

¹ Business Information Systems, University of Oldenburg, 26111 Oldenburg, Germany
{hahn, hausmann}@wi-ol.de

² Research Center for Information Systems in Project and Innovation Networks
University of Osnabrueck, 49069 Osnabrück, Germany
frederik.ahlemann@ispri.de

³ BTC AG, 26121 Oldenburg, Germany
sabels@acm.org

⁴ OFFIS e. V., 26121 Oldenburg, Germany
jan.strickmann@offis.de

Abstract. This paper introduces “PROMONT”, a project management ontology. PROMONT models project management specifications from a number of sources, most notably the upcoming DIN 69901 model for the exchange of project data. As a reference ontology it helps to build a common understanding of project related terms and methods thus facilitating the management of projects conducted in dynamic virtual environments. It is especially well suited for cross-enterprise project related business processes such as integration management, communication and controlling. A distributed project scenario is used to illustrate how a project plan is broken down for individual partners, and how activities are coordinated in a heterogeneous group.

1 Introduction

Project management is of vital importance in any organization. Whether it is product development, re-organization, mergers or marketing activities – most innovative endeavors are carried out as a project. Projects have limited duration, are conducted by a specially assigned organization and tackle a new task, thus dealing with a certain degree of uncertainty and risk. To meet these special requirements and risks, numerous methods for project management have been devised and successfully employed. Contemporary project management software systems (PMS) implement these methods and facilitate project planning, execution and control in daily business. Surprisingly, however, there is no common international data standard for the structure and exchange of project data in a broad use. Microsoft Office Project’s Schema has established itself as a “de facto” standard for some of the basic project management concerns, but it lacks a number of features such as support of risk or change management, integration with other systems or document management. Other project management solutions lack in similar features. These deficits are sometimes made up by other solutions such as portals, project file servers and administrative processes, creating

additional overhead but still failing to provide a simple integrated overview and control of a project.

To address these issues in a single integrated standard, a group of German researchers and PM software vendors formed a workgroup under the umbrella of the German Association of Project Management (GPM) and drafted a set of documents that are currently reviewed by a number of domain experts. The result will be issued to the German Institute for Standardization (DIN e. V., [14]) in order to become a German national standard in 2007. The standard will not only contain a project management data model [1], but also comprises process descriptions, project management methods and basic definitions in the form of a glossary. International dissemination and advancement of the standard's definitions will be fostered by a non-profit interest group founded by the workgroup members.

Given the fact that the challenge of creating a common data standard is accomplished by the upcoming DIN standard and in order to provide a formal specification of its glossary, descriptions and processes, the authors of this paper introduce PROMONT, a **project management ontology**. Our aim in this paper is to discuss how daily work in a virtual project environment [11] benefits from semantic enhancement of standards in the project management domain.

2 Standards, Related Work and PROMONT

Project management is one of the key factors in successful projects (i.e. projects that attain their goal within their planned boundaries). Since projects tend to get more and more complex, project management is crucial. Without comprehensive project management efforts, most projects today would fail. But what exactly is project management?

2.1 Project Management Standards

The most prevalent project management document is the Project Management Body of Knowledge (PMBOK) provided by the Project Management Institute [18]. It states that "Project Management is the application of knowledge, skills tools, and techniques to project activities to meet project requirements." It identifies a comprehensive set of project management definitions which are "good practice" and "generally recognized" ([12], [6]). They are organized into five generic process groups and nine project management knowledge areas (shown in Table 1), which determine the scope of what is generally understood to be project management. Another well-known initiative is the International Project Management Association (IPMA) [19] focusing on the training and certification of project managers. Its standards are set forth in the International Competency Baseline (ICB) [10] These two standards set the scope of project management definitions, so any endeavour to standardize project management data must bear them in mind and we will use the PMBOK knowledge areas to determine the applicability of project management data models related to our approach.

Table 1. Project management process groups and knowledge areas as defined by the PMBOK [12].

Process Groups	Knowledge Areas
<ol style="list-style-type: none"> 1. Initiating Process Group 2. Planning Process Group 3. Executing Process Group 4. Monitoring and Controlling Process Group 5. Closing Process Group 	<ol style="list-style-type: none"> 1. Project Integration Management. 2. Project Scope Management, 3. Project Time Management, 4. Project Cost Management, 5. Project Quality Management, 6. Project Human Resource Management, 7. Project Communications Management, 8. Project Risk Management, and 9. Project Procurement Management.

2.2 Related Work

Today one can find several initiatives that aim at collecting project management knowledge in some kind of standardized data model which can be used to implement project management software and to exchange project data. In order to perform project management activities people use different methodologies according to their needs and standards. Instead of creating a project plan manually, companies use project management software that supports most important tasks that appear in project management processes. Those software solutions emerged in the 1980's and have unfolded their full potential in the last decade.

Office Project is one of the most often used solutions available today (see [15] and [9]). Although it is not based on an official standard, it can surely be considered as a de-facto standard because of its market position. However, Microsoft Office Project does not have an open structure but uses a proprietary data model which is not defined by an independent body. Furthermore, it only focuses on a small subset of what is typically understood as project management in general. So, for example, the Microsoft Office Project's XML schema (MSPDI, Microsoft Project Data Interchange) only addresses three out of the nine knowledge areas, namely scope management, time management and cost management. The remaining areas are not supported or only in small parts. It is therefore not sufficient to use it as the only reference when designing new project management software.

To provide an open standard for the exchange of project management data, PMXML has been designed by Pacific Edge Software in the year 2000 (see [3]). It has been revised in 2002 and implemented by a few project management systems (Pacific Edge and Primavera), but has failed to achieve wider acceptance so far ([13]).

The most recent endeavor to create an open standardized data model for project data exchange is the DIN workgroup mentioned earlier. The standard has been developed by a consortium of eleven companies developing project management software and will probably be published with the next revision of DIN 69901. The first version of this standard was published in the early 1970s and has been revised several times since then [5]. By providing a data model, extended process descriptions and an outline of basic project management methods, the new version of DIN 69901 goes far

beyond its earlier versions and explicitly considers PMXML and is compatible with MSPDI.

The data model intended to be part of the new DIN 69901 [1] aims at supporting a wide range of project information that is needed in project management scenarios. For example, in extension of MSPDI it also considers:

- Change Management
- Document Management
- Quality and Risk Management
- Procurement Management
- Human Resource Management

2.3 Ontology Based Approach

All of the above approaches are based on textual descriptions or UML-models, but lack formal definitions on how data is to be interpreted by project partners. This is crucial when it comes to *integration and communication management*, the most difficult and yet most important of the nine knowledge areas defined by the PMBOK. Hence the authors propose an ontology-based approach as an extension of the DIN 69901 data model – PROMONT. By covering all issues from the DIN 69901 data model *plus additional semantic information* it is a comprehensive data structure able to handle all requirements that:

1. Are covered by project management software (software relevance),
2. Are necessary to exchange project information among various heterogeneous information systems (data exchange relevance),
3. Are necessary to fulfill common project management requirements (project management relevance).
4. *Are necessary for heterogeneous organizations and systems to successfully work together (interoperability relevance).*

When looking at existing research results that are similar to our efforts, the “Project Metrics Ontology” [16] and an ontology derived from the IT-CODE project [17] need to be mentioned.

- Project Metrics Ontology (PMO): This ontology was originally developed by BBN Technologies / Verizon in 2002 and is not meant to cover all typical project management issues. Instead of this, it focuses on providing an ontology that represents metrics for a specific project. This allows for example to perform performance measurements of specific projects or sub-projects.
- IT-CODE ontology: An ontology was defined to describe the project team of a building project within the IC-CODE project itself [17]. This ontology focuses on providing typical project management classes such as “Task”, “Actor”, “Project” or “Activity”. It contains 32 different classes and an additional set of 26 properties.

The PMO is a rather simple and short ontology containing only two classes and five properties. It is therefore not suitable for many project management tasks but it

can help to offer a reference when specifying project metrics. The second ontology covers a much broader scope compared to the PMO ontology but it focuses on building projects only and is not capable of representing all major, important concepts in the DIN specification which are needed for distributed project coordination. For example, it lacks risk management or milestone-concepts.

The features of PMO and IT-CODE do not fulfill all knowledge areas of the PMBOK, however they were taken into consideration in the definition of PROMONT.

3 Definition of a Project Management Ontology

While a standardized data model is good for the exchange of project data between PM-Software, it can also help to build a common understanding of terms and definitions in the field of project management, thus fostering interoperability not only for the exchange of data but also on a business process and organizational level. The DIN-data model is well suited for the implementation of PM-Software, since the UML-definitions can easily be used as a system reference. But, for the purpose of representing project management knowledge and project controlling, a project management ontology is a valuable extension for the representation of project management data since it is semantically more powerful than a data model with explanatory text. As a “formal, explicit specification of a common conceptualization” [7] an ontology provides means for expressively stating axioms and specifications of the concepts and relations in a certain field of expertise. Ontologies are both human and machine readable, abstracting from implementation technologies, data structures, system architectures, or applications (e. g. [2]). The core elements of the proposed project management ontology are described in the following sections.

3.1 Core Concepts of PROMONT

PROMONT formalizes the typical elements used for project structuring (such as task, milestone, resource or checklist). This formalization supports computer-aided project planning and evaluation to improve management decisions in a control circuit for project management.

The first step is to create a concept hierarchy which puts terms from the field of project management into a “sub concept-of” relationship, thus semantically refining the expressions. For example, the set of activities is defined as a subset of tasks, the crucial criterion being the assignment of a specific resource to a task. So any task with an assigned resource is also an activity that can be planned accordingly. These formal semantic rules narrow down the meaning of terms enhancing the expressiveness of the ontology [4].

Table 2 gives an overview of the core concepts of PROMONT. The semantics of each term are also defined. This excerpt from the taxonomy together with a few generic concepts forms the core of PROMONT. It can be expanded with domain- or method-specific terms to reflect a formal framework for project management and controlling ([See also [8]).

Table 2. Core concepts of PROMONT

Concept	Definition
Initiative	Any intention or endeavor, super ordinate concept for project, task and process.
Project	A project is a structured approach to deliver a certain result. It consists of an amount of time, budget and resource restrictions and conditions and is usually divided into a set of tasks.
Task	Project-specific initiative. May be divided into sub-tasks. May be implemented by the application of a process.
Activity	Task that has been assigned to a specific resource. The assignment determines duration and costs of task execution.
Phase	Subdivision of Project timing with specific objective. Often ends with a gate.
Resource	Consumable or not-consumable good or entity. Resources are necessary to execute an initiative.
Employee	Person working for an organizational unit. Subset of Resource.
Machine	A non-consumable, non-human Resource.
Calendar	A timetable showing the availability and workload of a resource during a period. Also used for a project overview calendar.
Skill	Property and potential of a resource to satisfy a requirement for a task.
Event	Occurrence of an action at a specific point in time. Has zero duration and may trigger tasks.
Milestone	Event with significant meaning for project status.
Gate	Milestone ending a phase. Usually associated with a formal review task.
Risk	Possible source of shortcomings or failure in the project. Might be sanctioned.
Objective	Desired outcome of an Initiative. Can be a physical product, a service or a document.
Result	Actual outcome of an Initiative. Can be a physical product, a service or a document.

3.2 Core Relations of PROMONT

Relationships between the above concepts are categorized and discussed in Table 3. The categorization builds on the fact that, just like concepts, relations can be arranged into a “is derived from” hierarchy. A derived relation refines the semantics of its more general super-relation, but still can be interpreted by an algorithm that does not know the specific derived relation.

3.3 Advantages of Ontology-Based Project Management

Having an ontology as a common, technology-independent, yet machine-readable exchange format for project management data will reduce the amount of interfaces and

Table 3. PROMONT relation categories and examples

category	Definition/example	derivations
attends	Shows, that the subject instance attends to the target instance. For example, a worker is responsible for a certain machine. Usually requires with a certain ability of the subject.	assigned to, responsible for
depends on	Shows the logical dependency of the target instance from the subject instance. Works usually depending on status, for example makes the start of a certain activity dependant on the conclusion of another. Inverse relations are <i>implies</i> and <i>supplies</i> .	needs completion of, hasPredecessor
implies	Models the logical consequence of the target instance from the start instance. The use of a resource implies certain costs, for example. Often inverse relation to <i>depend on</i> .	costs, triggers, needs
part of	The subject instance is a member, content or a component of the target instance. A human resource, for example, is a member of an organizational unit.	works for, is subtask, works for project
supplies	The subject instance offers a property of ability for the target instance, e. g. a Human Resource supplies his knowledge of engineering to a task. Inverse relation to <i>depend on</i> .	has ability, satisfied by
startsWith	Relation between initiatives and events to illustrate that an event triggers an initiative	-
endsWith	Relation between initiatives and events to illustrate that an event is triggered by an initiative.	-

mapping tables between different project management systems implementing the DIN 69901 or any other proprietary data model. The formal semantic model provided by PROMONT brings project data from sigmatic level to a semantic net which not only helps to exchange pieces of information but also provides their exact meaning. Established as a semantic net, PROMONT can be easily extended to cover a project’s context by merging with a domain specific ontology. Apart from easy data handling and expandability, the added semantic expressiveness of one integrated model is a major advantage of PROMONT. By applying first order logic and inference mechanisms, knowledge can be extracted from ontology instances, thus allowing better project control and improved decisions in the field of application as shown in the following example.

4 Communication and Integration Management Using PROMONT in a Virtual Project Environment

As stated above, communication and integration management are the knowledge areas of project management that would benefit the most from an “ontologization” of

PM-data models. In this section we show two examples how communication may be improved and the integration process controlled with the help of PROMONT.

Communication is a key success factor in any project, even more so in a virtual project community that can for example be found in cross-enterprise projects. PROMONT summarizes all important terms of project management and their exact semantic meaning. It provides project participants with a complete vocabulary to resolve ambiguous issues in the interpretation of project plans and elements. For example, most project participants define a milestone as an event with significant impact on the project status, often reached with the completion of a task or a sub-goal. Some PM-Tools however depict milestones as special tasks with a duration of zero or even identify them with a milestone flag. A commonplace example is shown in Table 4, where a task is identified as a milestone in MSPDI-format with the <Milestone>-tag. Using an XSL-Transformation it is converted into two separate concepts of PROMONT: a task associated with a milestone, thus resolving any possible ambiguity.

Table 4. Excerpt from two project plans in MSPDI and OWL-format. A task shown as a milestone in the former is converted into a task and a milestone in the latter.

<pre> <Task> <UID>2</UID> <ID>2</ID> <Name>Review</Name> <Duration>PT40H0M0S</Duration> <Milestone>1</Milestone> <EarlyStart date="2006-07-07" /> <EarlyFinish date="2006-07-07" /> <LateStart date="2006-07-07" /> <LateFinish date="2006-07-13" /> </Task> </pre>	<pre> <Task rdf:ID="Task_16537"> <rdfs:label>Review</rdfs:label> <rdfs:comment>Evaluation of project progress and recent results. </rdfs:comment> </Task> <Milestone rdf:ID="Milestone_28837"> <rdfs:label>Milestone</rdfs:label> <rdfs:comment>Marks successful completion of this projectphase. </rdfs:comment> <reachedIfFinished rdf:resource="#Task_16537" /> </Milestone> </pre>
--	--

Similarly, PROMONT is able to enrich the semantics of the model in even more sophisticated terms by stating rules such as:

- No resource can be assigned a daily workload of more than 24 hours.
- Subprojects and tasks must start and end within the boundaries of their superstructure.
- The hierarchy of program, project, sub-project, work packages, tasks and subtasks shall be respected.

When using the ontology in the background, a potential software application will gain from the semantics encoded. For example, it will then be sufficient to create a new initiative, i.e. instantiating the most general concept available, and then run a reasoning algorithm (such as Pellet [20] or Racer [21]) to determine the exact type of the object at any given point in time. These options facilitate communications because they give clear interpretations of data otherwise not available.

Closely related to communication management is the task of integration management with its three core processes project planning, execution and control. Different project plans, management methods and data formats need to be integrated into one general project plan. The execution of this plan needs to be closely monitored and results need

to be formally reviewed and released, to make sure all partial results fit to each other and the overall project goal is accomplished. The scenario as depicted in Fig. 1 illustrates how concepts and relations of PROMONT help to solve integration tasks.

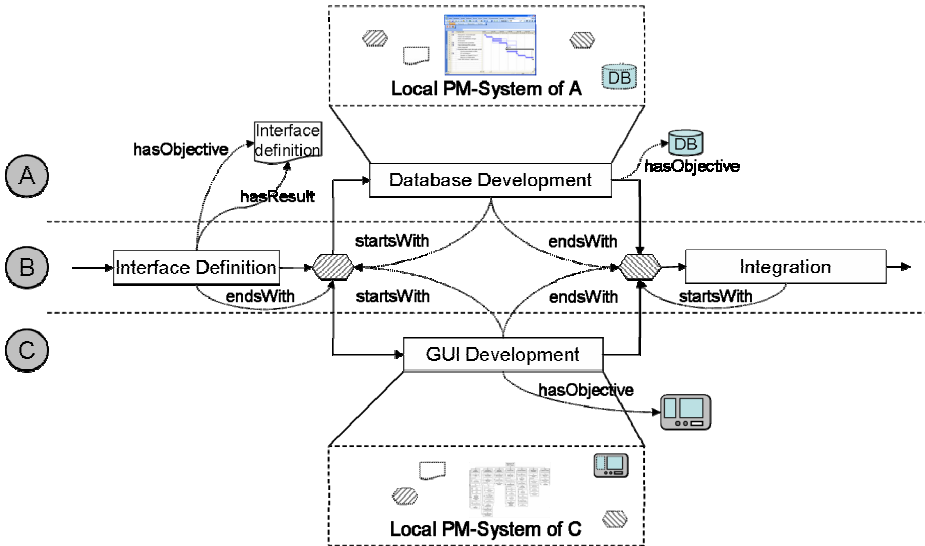


Fig. 1. A project plan logically structured by Events. Individual Tasks can be planned with local PM-Systems.

In this scenario, three project partners have to develop a Software system consisting of a database and a GUI. The overall project plan as designed by project partner B consists of the tasks “Interface Definition”, “Database Development”, “GUI Development” and “Integration” where the two development tasks are carried out by specialists (A, C). The Project plan defines subtasks with individual partial results (the document “interface definition” and the products “Database” and “GUI”) that in total shall lead to the desired output. PROMONT models these with the concepts “Objective” and “Result”, both with Document, Service or Product as sub-concepts. They allow specifying a set of desired or actual outcomes for any Initiative. This facilitates the analysis and integration of project results and helps to identify deviations once a project result has been delivered.

Another new and important concept provided by PROMONT is “Event” with the Relations “startsWith” and “endsWith”. They connect Initiatives to Events, a fact usually not explicitly stated in project plans. PROMONT uses Events for two purposes:

1. Events and their more specific sub-concepts Milestone and Gate can be used as cross-organizational reference points for the project status, without having to know in detail the initiatives and tasks that lead to the event or are triggered by it.
2. An Event can be used to trigger an automated process which implements a Task.

The Event-concept helps to analyze and integrate project tasks since events can be communicated as “bearings” to any project partner. As long as he delivers the defined results within these coordinates, he is free to structure his sub-project as he wishes. In the example scenario, the existence of the “interface definition” via “has result” triggers the event by “endsWith”. From this point onwards, the two individual project plans of A and C can be carried out without the need to share detailed information of their work. As soon as the two Objectives “DB” and “GUI” are fulfilled by fitting results, the next event can trigger the task “Integration”.

5 Conclusion and Further Research

In this paper, we have introduced PROMONT, an approach of defining a project management ontology that is based on the ideas of the upcoming DIN 69901 norm. PROMONT does not mean to replace the DIN approach, but it summarizes all major project management standards and tools in one integrated reference model. It offers extending definitions of project management issues aimed at supporting interoperability of project management systems, processes and organizations. In particular, PROMONT offers a formal approach to define relationships and conditions between different terms that are used in project management. Thus it is well suited for Communication and Integration Management the most crucial issues in distributed virtual projects. An example has been given to illustrate how events are used to structure a project into independent sub-projects coordinated via Events. It also shows how individual objectives can be defined, communicated and compared with actual results.

The future course of research will consist of the following steps:

1. Completion of PROMONT to cover full content of the DIN 69901-data model
2. Evaluate project management methods and processes to show how they complement with PROMONT
3. Implement a prototype system using PROMONT in a real project environment, using existing PM-tools as data sources to verify interoperability

To follow the steps outlined, PROMONT is currently being used in the PERMETER project aimed at measuring the performance of development projects. PROMONT is available for download on the PERMETER-website (www.permeter.de). First results can be expected in October 2006. Afterwards, we will start with a detailed evaluation of our ontology. We intend to demonstrate how to “convert” a real-world project into the PROMONT ontology and we will demonstrate the benefits that can be reached with PROMONT. Furthermore, we will compare different metrics in order to measure the practical applicability of PROMONT.

The Event concept helps to analyze and integrate project plans from various partners, the Objective/Result-concepts provide detailed possibilities to define required outputs and compare them with individual results. This will show its advantages especially once PROMONT is semantically integrated with other domain ontologies (e. g. for mechanical or software engineering) because then outcomes and results can be described in even more detail. Thus, PROMONT allows a much greater degree of interoperability both between PM-Systems and with other application systems in the project domain.

References

1. Ahlemann, F.: GPM Fachgruppe "DIN 69901 Neu", DIN-PM-Datenmodell - Erläuterungen, DIN 69901 specification draft, 2006.
2. Angele, J.; Nierlich, A.: Semantic Business Integration – Speed Up your Processes, http://www.ontoprise.de/content/e5/e69/e206/SemanticBusinessIntegration_ger.pdf, last access: 03/31/2006.
3. Curran, K.; Flanagan, L.; Callan, M.: PMXML: An XML Vocabulary Intended for the Exchange of Task Planning and Tracking Information. In: *Information Technology Journal* 3 (2): pp. 192-195, 2004.
4. Daconta, M. C.; Obrst, L. J.; Smith, K. T.: *The Semantic Web*, Wiley, 2003.
5. Deutsches Institut für Normung, DIN 69901, *Projektwirtschaft; Projektmanagement; Begriffe*, Beuth Verlag, 1987.
6. Duncan, W. R.; *A Guide to the Project Management Body of Knowledge*, PMI Standards Committee, Project Management Institute, 1996.
7. Gruber, T.R.: A translation approach to portable ontologies. In: *Knowledge Acquisition*, 5(2), 1993: pp. 199-220.
8. Hahn, A. Hausmann, K., Preis, S., Strickmann, J.: Ein Konzept für das Entwicklungscontrolling auf PLM-Basis. *HMD* 249, June 2006.
9. Meyer, M. M.: *Stand und Trend von Softwareunterstützung für Projektmanagement-Aufgaben*. Universität Bremen, GPMA, Bremen, 2005.
10. Motzel, E. et. al.: *ICB - IPMA Competence Baseline*, International Project Management Association, 2nd ed., 1999.
11. Munkvold, B. E.; Evaristo, R.: "Collaborative Infrastructure Formation in Virtual Projects," in H. M. Chung, (ed.) *Proceedings of the Sixth America's Conference on Information Systems*, August 10-13, 2000, Long Beach, California, USA: Omnipress, pp.1705-1710.
12. Project Management Institute: *A Guide to the Project Management Body of Knowledge (PMBOK)*, Project Management Institute, 2004.
13. Volz, R.: *PMXML - a XML standard for project management*, 2002, http://www.vrtprij.com/content/istandards/pmxml_en.html , last access 07/10/2006.

Web References

14. DIN: <http://www.din.de>
15. MS-Project: <http://office.microsoft.com/project>
16. PMO ontology: <http://www.daml.org/ontologies/349>
17. IT-CODE project: <http://www.civil.auc.dk/~i6ycl/itcode>
18. Project Management Institute: <http://www.pmi.org>
19. International Project Management Association: <http://www.ipma.ch>
20. Pellet OWL reasoner: <http://www.mindswap.org/2003/pellet/>
21. Racer OWL reasoner: <http://www.racer-systems.com/>

SPI Methodology for Virtual Organizations

Paula Ventura Martins¹ and Alberto Rodrigues da Silva²

¹ INESC-ID, CSI/Universidade do Algarve
Campus de Gambelas, Faro, Portugal
pventura@ualg.pt

² INESC-ID /Instituto Superior Técnico
Rua Alves Redol, n° 9 –1000-029 Lisboa, Portugal
alberto.silva@acm.org

Abstract. This paper discusses the importance of software process improvement in a virtual environment where several organizations are cooperatively involved in the development of a software product, each one using its own development process. The main focus of the paper is a methodology, called Process and Project Alignment Methodology, to improve the development process of a single organization based on projects knowledge. However, the authors believe that the same fundamentals can be applied in a virtual organization and discuss the extension of the presented methodology to a virtual organizational context.

1 Introduction

New and emerging market conditions are the core engines driving organizations in focusing on their competences and to cooperate with others under networks called Virtual Organization (VO), where each member has its own organizational culture and, in the context of software development, each one has its specific development process. Therefore, it is necessary to create a common specification to represent different processes for each organization. After this unified process representation, it is possible to define VOs development processes. To do so, does not mean to represent in a model every detail of all organizational processes involved in the VO. Instead, the organizations that participate in the interactions only have to describe interface issues. Software development process details which are internal to an organization should not be represented at this level but must be properly encapsulated. However, the involved context requires each organization to have an approach to process management before dealing with VOs development processes.

For now, our focus is to create a methodology to define and improve an organizational development process. Since project management is the discipline that controls and monitors deviations from the original project plan and also manages all process disciplines, project management is the right way to detect changes in the project that can lead to process improvement. Considering the relationship between the process and the project, new software process improvement (SPI) approaches have to consider process and project alignment and iterative SPI performed by project teams. Currently, there seems to be a lack of support on how SPI approaches

addresses the problematic about how development processes are effectively applied and improved using knowledge from software projects. Another challenge is how to control and validate important project changes that must be integrated in the process. In this paper we propose a SPI methodology based on process and project alignment that enables improvements on organizational processes. We also discuss and propose a solution that enables interactions between a virtual project and projects locally executed in different organizations. This methodology is supported by two meta-models that allow project definition based on a software process. We propose to extend these meta-models by allowing publishing and subscribing to events, and by enabling the definition of points in the project where events should be sent or received.

This paper is organized in the following sections. Section 2 presents a literature overview about alternative approaches to project management, process management and SPI. In the context of VO, we will present a brief description about VO projects related with process and project management. Section 3 discusses the problematic about SPI, process and project management. Section 4 describes the proposed methodology to support iterative SPI based on process and project alignment. In this section, we also discuss mechanisms to extend meta-models of the proposed methodology to support SPI in VOs. Finally, Section 5 presents conclusions, limitations of the methodology and future work.

2 Related Work

Process and project management is discussed by Budlong, Szulewski and Ganska [1], Climitile and Visaggio [2] and Chan and Chung [3]. But, only the AHEAD approach [4] has a fundamental feature: its support for process improvement. However this SPI solution isn't implemented in a project management context.

Considering that process and project alignment conduce to SPI activities, we present two SPI approaches discussed in the literature. Traditional SPI methods and approaches are based on final project retrospectives [5]. In these methods, there is a long time span between the problem identification and the validation of the new process. On the other hand, agile SPI approaches have a different perspective. According to agile principles [6], the project has reflections meetings in regular intervals. Cockburn proposes a reflection workshop technique [7], Dingsøy and Hanssen have a workshop technique called postmortem review [8], whereas Salo and Abrahamsson discuss a Post Iteration Workshop (PIW) method [9].

Therefore, all these approaches have no solutions to provide project management based on a process description and also iterative SPI. The main challenge in iterative SPI is to have project changes in real time, so project management must include fast feedback from each member of the project team.

Since this paper subject is about SPI in VOs, we present a short review about interactions between processes in VOs. Various forms of process interactions types are defined in literature, which we briefly summarize: capacity sharing, chained execution, subcontracting, (extended) case transfer, loosely coupled, public to private approach [10]. The problem of process management in VOs also had been addressed

by approaches using the notation of agreements and contracts, like the WISE [11] and CrossFlow [12] projects. However, these approaches do not present any concrete process management model. A detailed and interesting approach to process management in VOs has been proposed in the context of CMI project [13]. In the meantime, the problematic about effective use of the development process in virtual projects persists.

The core foundation of this paper is on how VOs manage their processes and keep projects aligned with that processes. So we describe two projects related to process and project management: (1) Intelligent Services and Tools for Concurrent Engineering (ISTforCE) and (2) Global Engineering and Manufacturing in Enterprise Networks (GLOBEMEN). ISTforCE is a European framework project, with the objective of designing a Web-based services platform through which engineers at a given design or consulting company will access the services on the Internet and collaborate in real time. It aims at creating infrastructure on which real construction companies and virtual teams of construction companies can rent and customize services on a project by project basis, and where providers of engineering services can market their products. In the ISTforCE, the authors stated that an Internet desktop system for engineers should have the following five requirements: it should be (1) open enough to integrate with other service or tools, (2) customizable to persons, (3) customizable to projects, (4) scalable, and (5) extendable [14]. Another project is the GLOBEMEN project, which aims to create IT infrastructures and related tools to support globally distributed product life cycle management, project and manufacturing management in the VO. The project focus is on VO information exchange and control on three core business processes of manufacturing industries: (1) interaction with customers and users including global product life cycle management, (2) optimization of the delivery chain through VO resource planning and (3) distributed concurrent engineering [15].

3 Problem Description

Project management, process management and SPI are interrelated disciplines that contribute to successful projects. So process management and SPI must be present during the entire execution of the project, even in the initial planning. Project planning is the most important phase in project management. The effort spent in identifying the proper needs and structure for organizing and managing a project could be minimized if the initial plan is process-based. Also important is the fact that project management monitors and controls activities from all the other process disciplines, so changes in these disciplines best practices will be detected throughout the project life cycle. SPI must be performed during project time and not only in dedicated evaluation periods. Projects are dynamic systems whose associated processes must always be under improvement.

However, many organizations have their development processes described but they don't effectively apply them in their projects. The defined process is not directly matched to their projects entities because organizations don't use process knowledge in project management.

4 Process and Project Alignment Methodology

Our research proposes a methodology that allows the definition, evaluation and improvement of an organization software development process. This proposal, called a Process and Project Alignment Methodology (PPAM), allows a general vision on the current state of an organization development process, as well as project alignment with the development process. Considering the theories and concepts described in the proposed methodology, we also discuss the mechanisms necessary to use PPAM in the context of VOs.

PPAM is based in a modelling approach since process and project modelling are the techniques used to define and analyze the significant aspects of development processes and projects. The proposed architecture identifies and interrelates the concepts necessary to provide SPI based on process and project management issues. In this paper we just show the application of the meta-models used to define processes and projects, a more detailed description of the meta-models is presented in [16]. This paper focuses on software process improvement and on how these meta-models can be used to solve this problem.

This section describes the components of PPAM essential to have process and project alignment. Process and project alignment formalization has four components: (1) process modelling enables an easy way to graphically construct a process; (2) project modelling (based on a process) provides the necessary coordination facilities for process and project alignment; (3) project control and monitoring enables observing changes in the project that are candidates to SPI. Process versioning enables creating process versions based on the proposed process improvement; and (4) process assessment allows the evaluation of the benefits due to process improvements.

Considering VOs, we propose a novel approach to virtual process and projects. This is supported by the idea that a virtual process or virtual project can be considered as a cooperation of several existing process or projects of collaborative organizations. The approach is inspired by the Service Oriented Architecture. Accordingly, the proposed methodology can be used in VOs environments but under some transformations in the following subjects: (1) virtual process definition; (2) virtual project creation and (3) groups involved. We discuss these key points as we present the components of PPAM.

4.1 Process Definition Component

A process meta-model provides a set of generic concepts to describe any process. ProjectIT Process Meta-model (PIT-ProcessM) architecture defines the concepts that correspond to elementary process concepts, allowing process creation or modification [16]. Two complementary views show those static and dynamic process elements. In the static view are represented the concepts related to process disciplines, like products, activities and roles. Meta-model dynamic view is about how a process life cycle is organized, e.g., phases and iterations. Additionally, all the process elements should be associated to a particular moment of a process life cycle.

At organizational level, the organization uses PIT-ProcessM to create his process. This step requires that the organization has knowledge about his practices based on historical data or from other process management initiatives.

In the context of VOs, PIT-ProcessM has to be extended with an interface to send and receive events. The processes created as instances from PIT-ProcessM will have an interoperability layer that identifies the services provided by each organization. The virtual process must be specified considering the services provided by collaborative organizations. Each individual organization must define its own supporting development process and must agree with other organizations on approaches to interoperability.

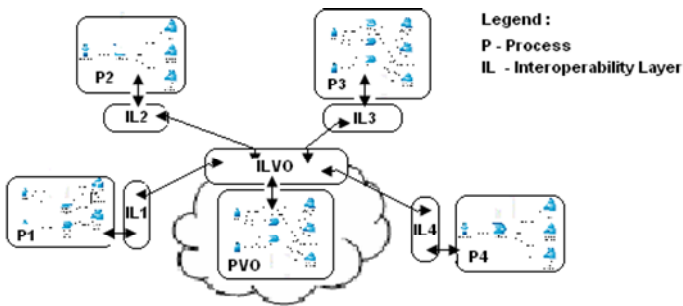


Fig. 1. VO Software Development Process

Fig. 1 presents an overview of the architecture for virtual processes. Each organization presents its own process (P1, P2, P3 and P4). The interoperability layer of each organization presents an interface to its process. The VO development process (PVO) will be defined considering the base features provided in each organization interoperability layer.

4.2 Project Definition Component

The second component (project definition considering a base process model) is essential to begin the project and consists in the project plan definition. A project is instantiated from a process, where a process represents reusable process practices at an abstract level. But in real world projects, multiple projects share the same process and are differentiated based on their specific elements, e.g. persons and the resulting relationships. Considering these differences and the process and project alignment carried out in this phase, our approach includes a ProjectIT Project Meta-model (PIT-ProjectM) to support this dependency [16].

Each organization defines their projects in alignment with a base process. In VOs, the enactment of the virtual project requires that each organization define its project and activate its interoperability layer in order to send and receive events. Event notifications will be managed by a suitable event service that is capable of filtering and correlating events, and of dispatching them to the appropriate target project.

The virtual project also needs a virtual team responsible by project activities and work products. A coordination organization must manage the virtual project. This

organization must create and delegate responsibilities for the virtual project control, monitoring and evaluation. Considering that virtual project change can improve the virtual process, some members of the participant organizations must perform the roles responsible by the SPI project.

4.3 Project Coordination and Monitoring Component

The third component consists in project coordination and monitoring. Updates and extensions to the initial project plan will be registered, always considering a base process model. Considering that some changes in the project best practices can improve the base process, we introduce a process versioning meta-model. The concept of process version has the advantage that the history of evolution of the processes is recorded. Since an effective process validation is only possible after a testing period, we must provide a mechanism to keep track of the changes carried out in the process's best practices. Thus, it isn't sufficient to maintain only the current version of a process. In this context, SPI subsumes two problems: (1) process modification and (2) ensuring that projects and base process remain consistent with each other.

Versioning Process. As proposed by agile methods, SPI is an iterative initiative during the project lifetime. Our proposal includes a workshop, when a dedicated member (process group) detects changes in project best practices that are considered as candidate improvements to the process. The basic idea of the proposed methodology is not to update process in place, but to version them. When a new process is created, this is considered the creation of a first version (root version). New versions are derived from existing ones by applying one or more modification operations to the based process version. However, as we will see, versions are created in an incremental way. Therefore, we will use the concept of versions states as used by [17], three states are distinguished: transient, released and obsolete. When a root version is created, it is in transient state. In this state, a version can be updated or deleted. In order to prevent invalid processes, when a version is in state transient its not allowed to: (1) create descendent versions; (2) create projects based on that version and (4) reference the version by another version. Finally, when a version is accepted its state is changed to released. In a released state, the version can't be deleted or updated, but all the other operations are allowed. When a released version has to be modified, his state is changed to transient, but only in special conditions (no descendents versions, no projects are based on it and is not referenced by other versions). If a version becomes unused is state is changed to obsolete and it is allowed to create new projects based on that version. An obsolete version can be deleted only: if has no descendents, has no derived projects and isn't referenced by other versions. But first it has to change to transient state and then the version can be deleted.

This section presents some details about extensions to PIT-ProcessM in order to support process versioning. Fig. 2 presents the main constructs of the extended meta-model. A process includes a unique identifier (process name) and a process version tree. A process version defines a version number and it is either in state transient, released or obsolete. A process comprises one or more process versions that can be derived from another process version by applying one or more modification operations. The diagram illustrates the relationships between a process and his

versions. PIT-ProcessM was updated to include modification operations applied to a process in creating a new version. Original elements from PIT-ProcessM like Phase, Iteration, Discipline, Activity, Role and WorkProduct are replaced by its versions classes. Associations between original PIT-ProcessM concepts are now performed between their version elements. Thus, the original elements have an association to its correspondent version, since each element can be used in one or more process versions.

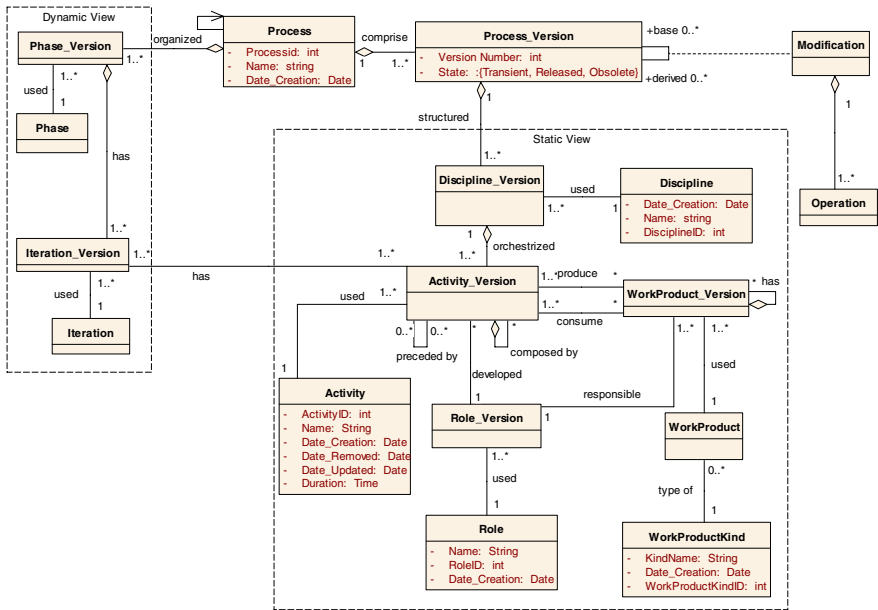


Fig. 2. PIT-ProcessM Versioning

In a virtual process, the versioning process will be performed in the same conditions as it was only one organization. However, the mechanisms used to introduce that changes can have two different sources: (1) changes in the virtual project practices or (2) changes in the processes of the participant organizations. In case of a change in the virtual project practice, the workshop proposed in PPAM has to be replaced by Internet technologies for supporting communication and collaboration (e-mail, audio or video calls, text chat, etc). But, if the new virtual process version is caused by changes in the services of involving organizations, the process group has to be responsible by the improvements in the virtual process.

Project Iterations. The groups involved in SPI consist on the software development team, project manager and the process group. The SPI method performed by these groups is realized throughout all iterations of a project, but the improvements follow a pattern that is performed in the time of two iterations (fig. 3). The SPI actions

performed in these two iterations are: (1) detect improvements and create a new process version (transient state); (2) test and validate the temporary changes in the next iteration of the project.

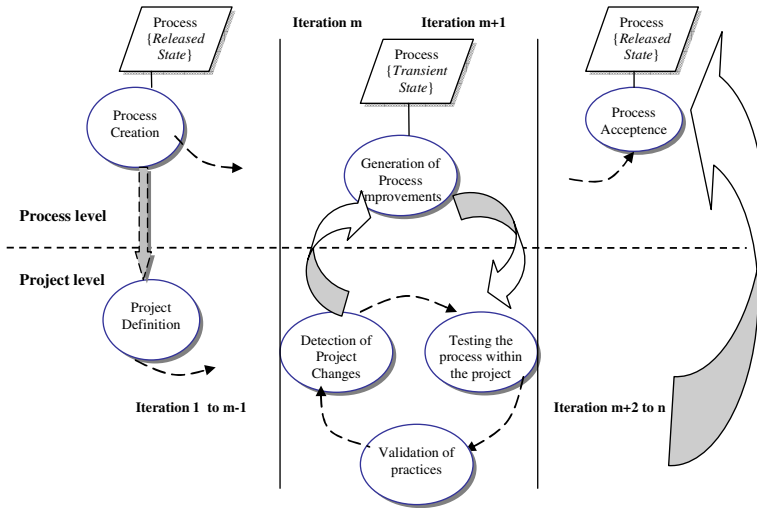


Fig. 3. Process and project alignment

In the first iteration, the project team must perform their daily work and detect situations that can lead to new practices in the project life cycle. The data collected includes positive and negative aspects found by the project team. The project manager has an important contribution in controlling the changes. At the end of this iteration, all candidate changes to improve the process are analyzed by the process group and, if necessary, a workshop is held to obtain more knowledge and present the improved process (new process version).

In the second iteration, team members get some feedback about the new practices and make notes to inform the process group. All this work will be under control of the process group. The project manager has to observe if team members are following the new proposed process. At the end of this iteration, all groups evaluate the work performed and decide if the process version is accepted. In case of success, the new process version is confirmed and the SPI method starts again. In case of failure, some new changes and improvements have been detected. The transient version will be updated and the evaluation work performed in the second iteration is repeated.

4.4 Process Improvement Assessment

In the final phase (SPI assessment), progress is evaluated throughout all process life cycle, specifying a set of improvements that can determine the process improvement itself. In the end of the project, process improvements must be analyzed in a reflection meeting. The main goal is to analyze all the improvement opportunities identified in the project and validate all the SPI actions accepted in workshops.

SPI activities can be performed successfully in many organizations. However, process managers ask themselves important questions following those activities, such as: (1) How to evaluate when a new process version meets the organization's goals?; (2) How to find if this is the appropriate development process?; (3) How will adjustments and changes affect the efficiency of the development process?; and (4) This new development process version will improve the performance of an organization?

Nowadays, project managers often lack reliable feedback from benefits of improving their development processes. In a daily basis, project managers use project management tools available in the market with the purpose of measuring and assessing software projects. However and due the fast changing environment proposed in the PPAM, the most important objectives are related with the impact on improving a software development process. Feedback information on SPI enables organizations to have control on future applications of a software process.

SPI assessment in practice can be viewed as the acquisition of data (key indicators) in a project where the new process version was applied, followed by data analysis and decisions about the further adoption of this development processes. Since project management is an important discipline in the proposed methodology, the key indicators must be those used by the project manager to analyse and evaluate a project. Normally, a project success is evaluated in terms of staff productivity, software quality, cycle time, and cost of the project. These features should be considered as key indicators to perform a SPI assessment.

5 Conclusions and Future Work

As organizations try to define their processes, they also recognize the need of continuous SPI. Even when the effort to improve them is done, organizations may fail at achieving the intended goals. Recognizing that the most critical problems occur during project activities, we strongly believe that process and project alignment can be a best-practice to get better project results and improve organizations software processes.

In this paper, we described the PPAM architecture and principles. The approach presents an innovative proposal but also includes ideas of other research initiatives in agile processes. We believe that PPAM contributes to a real process and project alignment. We present two meta-models that can support that alignment. The methodology concepts are being integrated in a project management tool (ProjectIT-Enterprise) of a research project (ProjectIT) from INESC-ID Information Systems Group [18]. In a VO domain, we discuss the extension of these meta-models to allow interactions between the virtual project and the projects executed locally in participating organizations.

As future work, our intention is to use the developed tool in real projects, to test and proof the approach. While, at the same time, new features could be found and included in the methodology. Additionally, these experiences will allow detecting new practices in the domain of organizational psychology, necessary to apply PPAM with success.

References

1. F. Budlong, P. Szulewski and R. Ganska, "Process Tailoring for Project Plan", The Process Management Technologies Team, The Software Technology Support Center (STSC), 1996
2. A. Climitile and G. Visaggio, "Managing Software Projects by Structured Project Planning", International Journal of Software Engineering and Knowledge Engineering, 7, 4, pp. 553-584, 1997
3. K. Chan and L. Chung, "Integrating process and project management for multi-site software development", Annals of Software Engineering, vol. 14, pp. 115-142, 2002
4. M. Heller, A. Schleicher and B. Westfechtel, "A Management System for Evolving Development Processes", Proceedings 7th International Conference on Integrated Design and Process Technology (IDPT 2003), Austin, Texas 2003
5. N. L. Kerth, *Project Retrospectives: A Handbook for Team Reviews*, Dorset House Publishing, April 2001
6. K. Beck et al., "Manifesto for Agile Software Development", <http://www.agilemanifesto.org/principles.html>, 2006
7. A. Cockburn, *Crystal Clear: a Human Powered Methodology for Small Teams*, Addison Wesley, November 2004
8. T. Dingsøy and G. K. Hanssen, "Extending Agile Methods: Postmortem Reviews as Extended Feedback", 4th International Workshop on Learning Software Organizations (LSO'02), Chicago, Illinois, USA, 2002, pp. 4-12
9. O. Salo and P. Abrahamsson, "Integrating Agile Software Development and Software Process Improvement: a Longitudinal Case Study", International Symposium on Empirical Software Engineering 2005 (ISESE 2005), Noosa Heads, Australia, 17-18 November 2005
10. W.M.P. van der Aalst, "Loosely coupled inter-organizational workflows: modeling and analyzing workflows crossing organizational boundaries", Information and Management 37, pp. 67-75, 2000
11. I.G. Alonso, C. Hagen, and A. Lazcano, "Processes in electronic commerce", in ICDCS Workshop on Electronic Commerce and Web-Based Applications (ICDCS 99), May 1999
12. P. Grefen, K. Aberer, Y. Hoffner and H. Ludwig, "CrossFlow: Cross-Organizational Workflow Management in Dynamic Virtual Enterprises", International Journal of Computer Systems Science & Engineering, Vol. 15, No. 5, pp. 277-290, 2000
13. D. Georgakopoulos, H. Shuster, A. Cichocki and D. Baker, "Managing process and service fusion in virtual enterprises", Information Systems, 24(6), pp. 429-456, September 1999
14. T. Cerovsek and Z. Turk, "Prototype Internet desktop for engineers", Product and process modelling in building and construction: proceedings of the third European Conference on Product and Process Modelling in the Building and Related Industries, Lisbon, Portugal, September 2000i
15. J. Laitinen, M. Ollus and M. Hannus, "Global Engineering and Manufacturing in Enterprise Networks GLOBEMEN", Proceedings of the Third European Conference on Product and Process Modelling in the building and related industries (ECPM2000), Lisbon, Portugal, September 2000
16. P. V. Martins and A. R. Silva, "PIT-P2M: ProjectIT Process and Project Meta-model", Proceedings of the OTM Workshop: MIOS+INTEROP 2005, Lecture Notes in Computer Science, Volume 3762, Agia Napa, Cyprus, pp. 516-525, October/November 2005
17. M. E. Loomis, "Object Versioning", Journal of Object-Oriented Programming, January 1992
18. A. R. Silva, "O programa de Investigação Project-IT", version 1.0, October 2004

A Pattern-Knowledge Base Supported Establishment of Inter-organizational Business Processes

A. Norta, M. Hendrix, and P. Grefen

Eindhoven University of Technology, Faculty of Technology Management, Department of Information Systems, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands
a.norta@tm.tue.nl

Abstract. In the domain of business-to-business (B2B) collaboration, companies are pursuing the objective of electronically linking their business processes for improving their supply chains. For creating such inter-organizational collaboration, intra- and inter-organizational knowledge workers (IKWs) function as assisting experts. However, IKWs must not constantly "reinvent the wheel" but should instead be supported by a repository that contains knowledge about how to design business processes. Thus, this paper proposes the support of IKWs by a pattern repository for the effective and efficient design of inter-organizational business processes. A pattern is conceptually formulated knowledge that is technology independent. By storing patterns in a uniform specification template of a meta model, it is possible to perform systematic reasoning. Having information readily available about the technology support of individual patterns, IKWs can quickly analyse with which intersection of pattern sets it is possible to link intra-organizational business processes.

1 Introduction

Companies that focus on their core competencies or miss know-how to perform certain business activities source services from providers. In this context, a service consists of a business process that is integrated into the in-house process of the consuming company. For example, a truck-producing company has suppliers of a water tank, an insurance company uses a third party to assess damage cases. A promising approach for B2B is the coupling of workflow concepts with service-oriented business integration. This emerging framework of dynamic inter-organizational business process management (DIBPM) [15] offers a new model for addressing the need of organizations for dynamically bringing together a service consumer and a service provider over web-based infrastructures where the service is a business process. To do so, DIBPM merges service-oriented business integration (SOBI) and workflow management concepts. The setup of such B2B commerce is a client-server relationship where one party offers a service that is integrated into the process of a consumer.

To establish intra- and inter-organizational business processes efficiently and effectively in DIBPM, the utilization of patterns is recommendable. Corporations typically comprise an information infrastructure consisting of a heterogenous system environment supporting their business processes. The situation turns even more complex when the business processes of collaborating parties are linked. By checking which patterns

the respective heterogeneous system environments support, a common denominator of collaboration is detected. Control-flow patterns [6,7,8] have been specified after investigating several intra-organizational workflow management systems. Furthermore, patterns for intra-organizational data-flow and resource management [21,22] have been discovered and specified. More recently so-called service-interaction patterns [9] have been specified for the coordination of collaborating processes that are distributed in different, combined web services.

In the domain of SOBI, web service composition languages (WSCL) have emerged for supporting process specifications, e.g., BPEL, BPML [10,11] and so on. Such languages compose services in a workflow, offering a complex service that carries out activities. The referenced pattern specifications and emerged WSCLs show that a rich amount of results exist that are relevant for DIBPM. For example, many e-business related patterns are textually available online [3] for the perspectives business-, integration-, composite-, custom design-, application- and runtime patterns. For inter- and intra-organizational knowledge workers (IKWs) who are exposed to business, technological, and conceptual complexity, such patterns promise a meaningful support for effectively and efficiently establishing inter-organizational business processes with the help of SOBI technology. IKWs organize the business processes in-house and establish business process links for B2B activities. They manage the heterogeneous system infrastructure that supports such business processes. However, the pattern specifications of various perspectives that IKWs need to employ, differ and it has not been investigated how they relate to each other across different perspectives.

Originating from the cognitive sciences, e.g., philosophy, psychology, the Language-Action-Perspective (LAP) facilitate the construction of automated, coherent messaging between information systems, as has been observed in many research works [13,16]. Briefly, LAP emphasizes what people *do* while communicating; how they create a common reality by means of language and how communication brings about a coordination of their activities. The approach of LAP is applicable in business collaboration [12] where inter-organizational transactions are intuitively modelled and carried out.

It is desirable to store all the pattern related data uniformly in one knowledge base and make it accessible for IKWs with tool support. Looking at the pattern repositories that are cited above, their content is always static and limited to either one or a couple of perspectives. However, for IKWs it is desirable to have a repository available that is interactive and dynamically growing in perspectives and content. The repository should store knowledge about how patterns relate to each other within the same perspective and across different perspectives. This paper fills the gap by proposing a pattern meta model that allows dynamic growth in content by permitting the admission of new patterns that may belong to newly introduced perspectives. Furthermore, a reference architecture is presented for the development of tools that use the pattern meta model and that support IKWs in employing patterns for the creation of intra- and inter-organizational business processes.

The structure of this paper is as follows. First, Section 2 describes how IKWs are involved in realizing inter-organizational business process collaboration. Next, Section 3 gives an overview of a pattern meta model that is used for uniformly storing and relating patterns to each other. In Section 4, the lifecycle of a pattern is used to deduct

requirements a pattern repository must fulfil that runs on top of the pattern meta model. Section 5 presents related work and Section 6 concludes the paper.

2 Inter-organizational Business Process Collaboration

This section gives a definition of DIBPM and relates inherent perspectives to each other. Furthermore, the nature of IKW involvement in DIBPM is made explicit. A definition of DIBPM [15] is given as follows: A dynamic inter-organizational business process is formed dynamically by the (automatic) integration of the subprocesses of the involved organizations. Here dynamically means that during process enactment collaborator organizations are found by searching business process market places and the subprocesses are integrated with the running process.

Important issues in connection with DIBPM are the definition and identification of processes, the way compatible business partners find each other efficiently, the dynamic establishment of inter-organizational processes, and the setup and coupling of inter-organizational processes for enactment. In Figure 1 different perspectives are depicted for creating an inter-organizational business process collaboration. To the left and right two factory symbols represent the collaborating organizations that have their internal legacy systems. Those legacy systems are linked with intra-organizational business processes that combine several perspectives. In Figure 1 the intra-organizational perspectives control-flow, resource, data-flow, and transaction are depicted. However, it is possible that further perspectives are included or omitted.

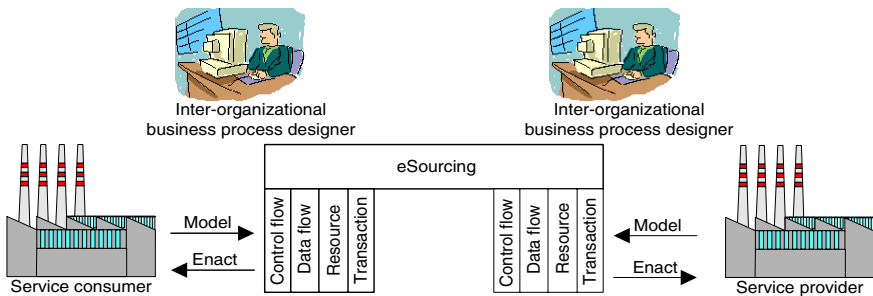


Fig. 1. Knowledge workers in inter-organizational collaboration

An additional perspective is contained in Figure 1 that crosses the boundaries of collaborating organizations, namely the eSourcing [17] perspective for which a catalogue of patterns [2,19,20] is specified. In the context of DIBPM, eSourcing is a framework for harmonizing on an external level the intra-organizational business processes of a service consuming and one or many service providing organizations into a B2B supply-chain collaboration. Important elements of eSourcing are the support of different visibility levels of corporate process details for the collaborating counterpart and flexible mechanisms for service monitoring and information exchange. For sake of brevity, [2]

contains in detail online eSourcing interaction and construction patterns, eSourcing-configuration examples, and corresponding documentation.

On top of Figure 1 two IKWs are depicted that each belong to a collaborating organization. Although the creation of eSourcing configurations should be carried out as automated as possible, it is realistic that IKWs remain provisionally necessary for the foreseeable future. For carrying out their work effectively and efficiently the support of a knowledge base is sensible that contains perspective-specific patterns for intra- and inter-organizational business process management. Thus, the next section presents the main building blocks of a meta model that is suitable for patterns of arbitrary perspectives.

3 The Pattern Meta Model

In the introduction of this paper several pattern sources of differing perspectives are referenced. The specifications of those patterns use templates with similar keywords. Many pattern specifications use a template of keywords that deviates or where the keywords are used differently. By harmonizing pattern-specification templates in a pattern meta model, IKWs can comprehend better the differences, commonalities, relationships of patterns. For sake of brevity, only the packages of the meta model are presented, while in [18] all classes and their relationships of the pattern meta model are depicted and explained.

3.1 Meta-model Packages

The left side of Figure 2 depicts a model of packages that are related to each other. These packages encapsulate classes that are explained in following sections. The center of the package-model is named `Pattern`, which contains all classes that capture information for specifying a patterns. In the `Taxonomy` package, classes are contained that capture information about DIBPM perspectives. This package contains classes that create a taxonomy into which patterns can be embedded. The `Support` package encapsulates classes for managing information about technologies that support patterns. Finally, the `User Management` package captures information of different users of the pattern repository, e.g., administrator, reviewer, pattern submitter, and so on.

On the right side of Figure 2 the core class of the `Pattern` package is depicted, which is equally named `Pattern`. The attributes of this class form the main description template of a pattern specification. A pattern has a `version` and a `name` that should be meaningful. Furthermore, a pattern has an `author` and a `creationDate` for every version. The `description` of a pattern mentions the inherent pattern properties and describes the relationship between them. Furthermore, the `intuitiveVisualization` contains a model that helps to support the comprehensibility of the pattern description. The `problem` of a pattern is a statement describing the context of pattern application. In this context conflicting environmental objectives and their constraints are described. The application of a pattern in that context should result in an alignment of the given objectives. Next, the `context` states a precondition, which is the initial configuration of a system before the pattern is applied to it. On the other hand, the `resultingContext` describes the postcondition and possible side-effects of pattern

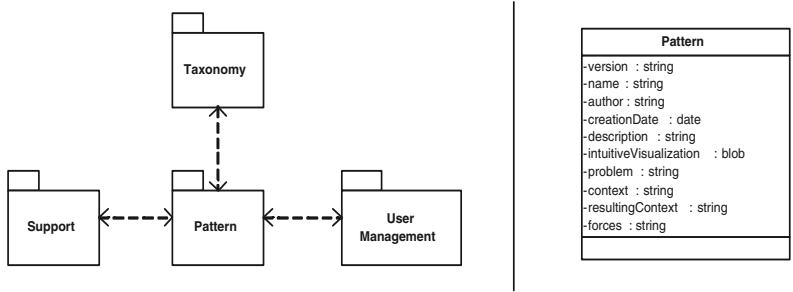


Fig. 2. Meta-model packages with their dependencies and the Pattern class

application. Finally, the *forces* describe trade-offs, goals and constraints, motivating factors and concerns for pattern application that may prevent reaching the described postcondition.

In Section 2 it is emphasized that IKWs need support by a pattern repository during creating inter-organizational business process collaboration. The next section explores the features of such an application.

4 An DIBPM Design Application

The lifecycle of a pattern is the starting point for deducting a reference architecture for a DIBPM design application. Briefly, in [18] it is explained that a repository user with the authorization of leading a review proposes the pattern for a review process. Repository users with the right skills may volunteer for a review or be explicitly invited by the review leader. Based on a defined review rule, a certain number of review results needs to be submitted for determining whether the pattern is accepted or not. If the review rule is not satisfied, the pattern is rejected and needs to be rewritten as a new proposal. If the review rule is satisfied, the pattern proposal is officially accepted and experiences a status change. Thus, it turns into a quality pattern that is exposed to IKWs for searching.

The pattern lifecycle is chosen as a starting point for deducting an extension of the pattern meta model. This extension is necessary for capturing additional information that is required for running an online application on top of the pattern meta model. The modules of the application architecture are deducted from the pattern lifecycle. Furthermore, the application architecture is also the result of experiences stemming from implementing a proof-of-concept prototype that is described in [18] together with screen shots.

4.1 An Application Architecture

An *author* is a user who submits a pattern to the repository. A *review leader* forms a review committee for the evaluation of newly submitted patterns. Registered users of the repository who indicate to be volunteers as *reviewers* are invited by the review leader to form a committee. An IKW with the role termed *analyst* is interested in browsing a

repository that contains patterns of different perspectives and corresponding information about their artifact support. As indicated in Figure 1, that pattern information helps an analyst to estimate which patterns collaborating business parties support despite their heterogeneous system environments. That way the setup time of inter-organizational business processes is accelerated. Finally, an *administrator* of the pattern repository is required to grant roles to registered users, troubleshoot during pattern reviews, and so on.

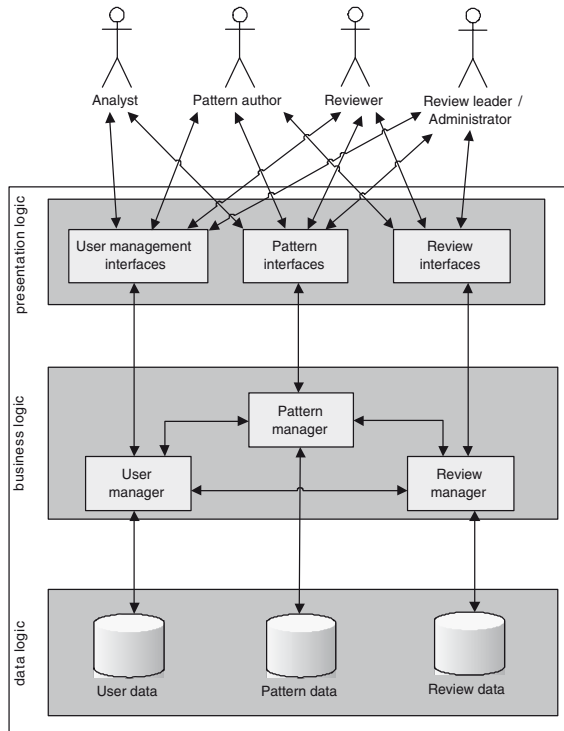


Fig. 3. The application architecture of the pattern repository

The mentioned roles for repository users are input for an architecture on top of the pattern meta model. The described repository user types are depicted in Figure 3 where bi-directional arrows indicate an exchange with certain modules of the application’s web interface. In the interface layer, modules are contained, for user management related interfaces, pattern related interfaces, and review related interfaces.

- The *user management interfaces* offer a repository user to register, claim various qualifications, and request particular roles. As different roles give a repository user different rights, the administrator may need to authorize the roles. Once a repository user is approved, the user management interfaces allow user to login and logout, and modify roles and qualifications.

- The *pattern interfaces* allow users to browse the repository for patterns with a search engine that uses facts from the classes belonging to the taxonomy package and the support package [18]. The generated lists of patterns can be individually selected for exploring their details.
- *Review interfaces* allow a review leader to set up a review committee consisting of reviewers who either volunteer or are appointed based on their qualifications. In the latter case an appointed reviewer may decline through an interface. After the reviewers explore the properties of a pattern, they submit an accept or reject and their feedback for the pattern author. The latter repository user checks the feedback from the reviewers through another interface.

The functionality layer of Figure 3 shows modules that support the web-interface layer, namely the *user manager*, *pattern manager*, and the *review manager*. They process input of the repository users and control the sequence of interfaces that are presented for the signing up and signing in of users, submitting and browsing patterns, performing reviews, and various administration activities. Figure 3 depicts the modules of the functionality layer are referencing each other. For example, to perform a review, the review-manager module uses functionality contained in the user-manager module. As a result competent review teams are organized with the right qualifications. During a review, functionality from the pattern manager allows a reviewer to explore the context a pattern proposal is embedded in, i.e., the taxonomy location, technology support, relationship with other patterns, and so on. Furthermore, a reference between the pattern-manager module and the user-manager module exists for the same reason of employing functionality from each other. For example, if a repository user wants to browse for pattern information, she needs to have the role of an analyst, which must be checked by using functionality from the user manager.

The bottom of Figure 3 depicts the data layer showing databases for *user data*, *pattern data*, and *review data*. These databases are referenced by the corresponding modules of the functionality layer. The figures of Section 3 contain data elements that are in the pattern data. The review and user data is explained in [18] without claiming completeness.

5 Related Work

Referencing patterns, Gamma et al. [14] first catalogued systematically some 23 design patterns that describe the smallest recurring interactions in object-oriented systems. Those patterns are formulated in a uniform specification template and grouped into categories. For the domain of intra-organizational business process collaboration patterns were discovered in various perspectives.

In the area of control flow, a set of patterns was generated [6,7,8] by investigating several intra-organizational workflow systems for commonalities. The resulting patterns are grouped into different categories. Basic patterns contain a sequence, basic splits and joins, and an exclusive split of parallel branches and their simple merge. Further patterns are grouped into the categories advanced branching and synchronization, structural patterns, patterns involving multiple instances, state-based patterns, and cancellation patterns. The resulting pattern catalog is for the evaluation [5,23] of WSCLs.

Following a similar approach as in the control-flow perspective, data-flow patterns [21] are grouped into various characteristics categories. One category is focuses on different visibility levels of data elements by various components of a workflow system. The category called data interaction focusses on the way in which data is communicated between active elements within a workflow. Next, data-transfer patterns focus on the way data elements are transferred between workflow components and additionally describe mechanisms for passing data elements across the interfaces of workflow components. Patterns for data-based routing deal with the way data elements can influence the control-flow perspective.

Patterns for the resource perspective [22] are aligned to a the lifecycle of a work item. A work item is created and either offered to a single or multiple resources. Alternatively a work item can be allocated to a single resource before it is started. Once a work item is started it can be temporarily suspended by a system or it may fail. Eventually a work item completes. The transitions between those life-cycle stages of a work item either involve a workflow system or a resource. Characteristic categories for the resource perspective are deducted from those life-cycle transitions and group specified patterns.

So-called service interaction patterns [9] are specified for the coordination of collaborating processes that are distributed in different, combined web services. Again, the patterns are categorized according to several dimensions. Based on the number of parties involved, an exchange between services is either bilateral or multilateral. The interaction between services is either of the nature single or multi transmission. Finally, if the bilateral interaction between services is of the nature two ways, a round-trip interaction means the receiver of a response must be equal to the sender. Alternatively a routed interaction takes place.

Many other e-business related patterns are textually available online [3] for the perspectives business-, integration-, composite-, custom design-, application- and runtime patterns. The business perspective highlights the most commonly observed interactions between users, businesses, and data. Integration patterns connect business patterns for creating composite patterns. Patterns of the composite perspective that combine business and integration patterns are only documented when they often occur in reality. The perspective custom design is similar to composite patterns. Finally, patterns from the application perspective focus on the partitioning of the application logic and data while patterns from the runtime perspective use nodes to group functional requirements that are interconnected to solve a business problem.

The INTEROP [4] network of excellence comprises a task group for methods, requirements and method engineering for interoperability. It is the objective of that task group to develop and validate ways of providing a knowledge repository of interoperable method engineering services. The pattern repository is to be integrated during the implementation of the method-chunk repository.

6 Conclusion

This paper proposes that intra and inter-organizational knowledge workers should employ patterns for dynamic inter-organizational business process management. Using

patterns promises the speedy evaluation and integration of intra-organizational business processes across the domains of collaborating parties. Since many patterns are specified in different perspectives for DIBPM, the need for a knowledge system in the form of a pattern repository arises to support IKWs. Thus, this paper describes a meta model for uniformly storing pattern specifications, orders them in a taxonomy, and caters for capturing information about technology support of specific patterns. An architecture for an online application is presented that builds on top of the pattern meta model.

With respect to ongoing research projects, the pattern repository is part of the proof-of-concept architecture for the EU project called CrossWork [1]. It is the objective in CrossWorks to develop automated mechanisms for allowing dynamic workflow formation and enactment, enabling hard collaboration and strong synergies between different organizations. Software agents in CrossWork employ a knowledge base for reasoning about automated workflow formation. Thus, by extending the pattern repository proposed in this paper with a formal second tier, agents can use facts about patterns for workflow formation. Furthermore, the pattern repository proposes itself as an integral part of a method-chunk repository that is built in the framework of the task group for methods, requirements and method engineering for interoperability of the INTEROP network of excellence.

Scope for future research exists for the repository prototype where the development of sophisticated graphical user interfaces constitutes a problem of considerable complexity. Furthermore, the prototype must comprise a powerful search engine that permits an inter-organizational knowledge worker to intuitively find suitable patterns with diverse combinations of data.

References

1. CrossWork: Cross-Organisational Workflow Formation and Enactment. <http://www.crosswork.info/>.
2. eSourcing: electronic Sourcing for business to business. <http://is.tm.tue.nl/research/eSourcing>.
3. IBM patterns for e-business. <http://www-128.ibm.com/developerworks/patterns/>.
4. INTEROP: Interoperability Research for Networked Enterprises Applications and Software. <http://interop-noe.org/>.
5. W.M.P. van der Aalst, M. Dumas, A.H.M. ter Hofstede, and P. Wohed. Pattern-Based Analysis of BPML (and WSCI). *QUT Technical report*, (FIT-TR-2002-05):487–531, 2002.
6. W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. Workflow Patterns Home Page. <http://www.workflowpatterns.com>.
7. W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. Advanced Workflow Patterns. In O. Etzion and P. Scheuermann, editors, *7th International Conference on Cooperative Information Systems (CoopIS 2000)*, volume 1901 of *Lecture Notes in Computer Science*, pages 18–29. Springer-Verlag, Berlin, 2000.
8. W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski, and A.P. Barros. Workflow Patterns. *Distributed and Parallel Databases*, 14(3):5–51, 2003.
9. A. Barros, M. Dumas, and A.H.M. ter Hofstede. Service interaction patterns. In W.M. P. van der Aalst and F. Curbera B. Benatallah, F. Casati, editors, *Business Process Management: 3rd International Conference, BPM 2005*, number 3649 in *Lecture Notes in Computer Science*, pages 302–318, Nancy, France, 2005. Springer Verlag, Berlin.

10. BPML.org. *Business Process Modeling Language (BPML) version 1.0*. Accessed August 2003 from www.bpml.org, 2003.
11. F. Curbera, Y. Goland, J. Klein, F. Leymann, D. Roller, S. Thatte, and S. Weerawarana. *Business Process Execution Language for Web-Services*. <http://www-106.ibm.com/developerworks/library/ws-bpel/>, 2003.
12. J.L.G. Dietz. The deep structure of business processes. *Communications of the ACM*, 49(5):58–64, 2006.
13. F. Flores, M. Graves, B. Hartfield, and T. Winograd. Computer systems and the design of organizational interaction. *ACM Transactions on Information Systems (TOIS)*, 6(2):153–172, 1988.
14. E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Professional Computing Series. Addison Wesley, Reading, MA, USA, 1995.
15. P. Grefen. Service-Oriented Support for Dynamic Interorganizational Business Process Management. to appear, 2006.
16. S. Kimbrough and S. Moore. On automated message processing in electronic commerce and work support systems: speech act theory and expressive felicity. *ACM Transactions on Information Systems (TOIS)*, 15(4):321–367, 1988.
17. A. Norta and P. Grefen. A Framework for Specifying Sourcing Collaborations. In Jan Ljungberg and Bo Dahlbom, editors, *14th European Conference on Information Systems: Grand Challenges*, pages CD-ROM, Gothenburg, Sweden, 2006. IT-University Gothenburg.
18. A. Norta and P. Grefen. A Pattern Repository for Establishing Inter-Organizational Business Processes. BETA Working Paper Series, WP 175, Eindhoven University of Technology, Eindhoven, 2006.
19. A. Norta and P. Grefen. Developing a Reference Architecture for Inter-Organizational Business Collaboration Setup Systems. BETA Working Paper Series, WP 170, Eindhoven University of Technology, Eindhoven, 2006.
20. A. Norta and P. Grefen. Discovering Patterns for Inter-Organizational Business Collaboration in a Top-Down Way. BETA Working Paper Series, WP 163, Eindhoven University of Technology, Eindhoven, 2006.
21. N. Russell, A.H.M. ter Hofstede, D. Edmond, and W.M.P. van der Aalst. Workflow Data Patterns. *QUT Technical report*, (FIT-TR-2004-01), 2004.
22. N. Russell, A.H.M. ter Hofstede, D. Edmond, and W.M.P. van der Aalst. Workflow Resource Patterns. *BETA Working Paper Series*, WP 127, Eindhoven University of Technology, Eindhoven, 2004.
23. P. Wohed, W.M.P. van der Aalst, M. Dumas, and A.H.M. ter Hofstede. Analysis of web services composition languages: The case of bpel4ws. In I.Y. Song, S.W. Liddle, T.W. Ling, and P. Scheuermann, editors, *22nd International Conference on Conceptual Modeling (ER 2003)*, number 2813 in Lecture Notes in Computer Science, pages 200–215, Chicago, Illinois, 2003. Springer Verlag, Berlin.

On the Concurrency of Inter-organizational Business Processes

Diogo R. Ferreira

Organizational Engineering Center, INESC and
IST - Technical University of Lisbon
Avenida Prof. Dr. Cavaco Silva, 2780-990 Porto Salvo, Portugal
diogo.ferreira@ist.utl.pt

Abstract. Within organizations, workflow systems can automate business processes by centrally coordinating activity sequences. But outside their borders, organizations are autonomous entities that cannot be subject to centralized process control. Their internal processes are autonomously defined and controlled, and what they need is to synchronize those concurrent processes. Just like Petri nets are a valuable tool to model activity sequencing in local business processes, π -calculus becomes a useful tool to model concurrency in inter-organizational processes. After a review of the main developments in cross-organizational workflow management, this paper illustrates the use of π -calculus to model the interactions between business processes running concurrently in different organizations. These interactions range from invoking external services to more complex patterns such as contract negotiation and partner search and selection. The paper concludes with a case study that illustrates the application of the proposed approach in a more realistic business scenario.

1 Background

Research on the application of workflow management in inter-organizational environments has produced a wealth of interesting approaches on how to deal with the problem of coordinating processes that span multiple organizations. The available solutions for cross-organizational workflow management have been developed in recent years and they have followed a path towards increasing flexibility, from focusing on “low-level” issues of workflow systems interoperability to more flexible, “higher-level” architectures based on contracts and views. Putting these developments in perspective, several trends can be identified, including:

Workflow interoperability mechanisms - after the publication of the Workflow Reference Model [1], which described four models of interoperability between workflow systems, supporting cross-organizational workflows seemed to be a matter of selecting the most appropriate interoperability model for a given scenario. In [2], Anzböck and Dustdar describe the application of these models to medical imaging workflows that resemble cross-organizational workflows.

In [3], van der Aalst formalizes the case-transfer and extended case-transfer models using Petri nets, and compares them as approaches to partitioning cross-organizational workflows over multiple business partners.

Federating heterogeneous workflow systems - it was quickly found out that run-time interoperability during process execution required build-time interoperability during process definition as well. The problem of connecting workflow systems then turned into a problem of federating them [4], i.e., to devise architectures in which several different workflow systems appear as a single, integrated one. Different solutions emerged, such as [5], in which Lindert and Deiters propose an approach to defining cross-organizational workflows via the interconnection of “process fragments” specified by different parties, in an early effort to address the problem of autonomy of business partners in inter-organizational settings. Reichert et al [6] realized the same problem, but focused on adaptive features that require the use of a centralized workflow modeling facility.

Agent- and service-based architectures - the paradigm of software agents renewed the interest in supporting cross-organizational workflows, especially in connection with service-oriented architectures, by facilitating the integration of local and remote services. Blake [7] developed a middleware architecture based on service-invoking agents which are essentially controlled by a global workflow manager agent that enforced a centralized workflow policy. Kwak et al [8] developed a middleware architecture where the workflow system inside an organization can invoke services registered either in a local service interface repository (LSIR) or in a global service interface repository (GSIR). Stricker et al [9] went even further to propose a trader system that promotes the reuse of workflow data types between organizations and supports bidding of service offerings to submitted service requests. Yan and Wang [10] devised an architecture where local processes are exposed as services that can be invoked in cross-organizational workflows.

Contract-based approaches - it had already been realized that the autonomy of business partners means that cross-organizational workflows are subject to agreements, or contracts. The CrossFlow project [11] was a milestone in the development of contract-based approaches, since it developed a consistent framework for establishing contracts and configuring workflow systems based on those contracts. Other authors have since then developed similar approaches, such as [12], in which van den Heuvel and Weigand propose a contract specification language to define cross-organizational workflows. But in [13], Kafeza et al realize that cross-organizational workflow contracts should refer only to partial views of local workflows. This is due to several reasons, including the need to keep some information private, or the fact that an organization may establish contracts with several partners having different requirements.

View-based models - the work on contract-based approaches opened up a new way of looking towards cross-organizational workflows. One of the most significant developments is [14], in which van der Aalst and Weske describe their public-to-private approach, where “public” stands for the agreed-upon workflow

and “private” stands for the local workflows running at each end. The private workflow can be obtained via inheritance [15] from the public workflow. Chebbi et al [16] describe how to do the opposite, i.e., how to obtain the cooperative (public) workflow from the local workflow, given the set of publicly advertised activities. In [17], Chiu et al present an XML-based language for describing public workflow views.

Apart from a few exceptions, which include for example the work of Chebbi et al [16] just mentioned, the centralized control of cross-organizational workflows - whether during build-time, run-time, or both - has been a recurrent assumption in cross-organizational workflow management. [4], [5], [6], [7], [9] are some examples of contributions that rely on the ability to centrally coordinate a cross-organizational workflow. [8] and [10] are examples of authors who escaped that problem by basing their approaches on activity outsourcing, hence giving legitimate control of the process to the contractor. Still, even in recent developments, such as process mediation by means of web service choreographies [18], authors often resort to centralized process control in order to coordinate cross-organizational workflows.

But inter-organizational processes require a different focus. Whereas activities and their sequencing are the main issue in workflow management within an organization, inter-organizational environments are dominated by interactions and concurrency. In these scenarios, organizations search for potential business partners, engage in contract negotiations, and establish channels in order to perform sets of interactions. Rather than enforcing an activity sequence, the need is to synchronize business processes running concurrently in different organizations. The purpose of this paper is to draw attention to the concurrent nature of inter-organizational processes and to show how one can resort to different techniques in order to model sequencing in one case, and concurrency in another. In this context, the main contenders for workflow modeling - Petri nets and π -calculus - can actually be used together in order to understand those two features of inter-organizational processes.

2 Modeling Concurrency with π -Calculus

Since the publication of Smith and Fingar’s paper [19], there has been an ongoing debate about the value of π -calculus for workflow management. Most of the controversy has been set around the ability of π -calculus to model certain workflow patterns, defined using Petri nets [20]. Recent work from Puhlmann and Weske [21] suggests that it is indeed possible to describe workflow patterns using π -calculus, although that is far from producing a workflow revolution, as originally proposed in [19]. Meanwhile, there are already some contributions on the application of π -calculus to workflow modeling [22] and to modeling the interactions of different entities in an electronic market [23]. Additional developments will certainly follow.

Apart from the controversy, π -calculus can be extremely useful to model concurrent business processes. In fact, π -calculus was devised having concurrency in mind, which makes it an obvious choice to represent synchronization points

between concurrent processes. On the other hand, we will keep on using Petri nets to represent activity sequencing for the local processes running within organizations.

The basic elements that we will be using to model inter-organizational processes are shown in figure 1. In this figure, *A* and *B* denote two autonomous organizations. Both of them have internal processes, and they will be interacting with each other while carrying out those local processes. The local processes are described using Petri nets, where each place is associated with an action, and each transition with an event¹. Each action stands for either a local task or an interaction with the external environment. In this case, only interaction tasks are shown. The interactions are represented using π -calculus links, the dot indicating the receiving end.

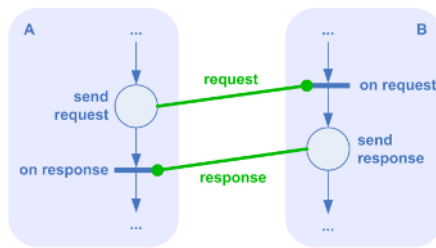


Fig. 1. Elements for modeling inter-organizational processes

For the sender, the interaction is an action that produces the outgoing message, so it is associated with a place. For the receiver, the interaction is an event that signals the arrival of that message, so it is associated with a transition. In the example shown in figure 1, after *B* receives the request from *A*, they swap roles; now it is *B* who produces an outgoing response, and *A* that receives the response as a transition-firing event. This example is illustrative of how an external service invocation - automated at both ends but not between them - could be modeled. In terms of π -calculus, this example could be written as:

$$A(link) = \overline{link} < request > .link(response) \tag{1}$$

$$B(link) = link(request).\overline{link} < response > \tag{2}$$

Basically, these two expressions specify what *A* and *B* are doing concurrently. *B* is given a link and waits for a request to be received through that link, then sends a response. *A* begins by sending the request through the given link, then waits for a response on the same link.

¹ At this point, it should be noted that the adopted approach follows [24], whereas some authors use transitions to represent workflow activities, as proposed in [25]. A discussion of these two approaches is beyond the scope of this article, but can be found in [26].

3 Modeling Contract Negotiation

In inter-organizational environments, where autonomous companies interact with each other, it is often difficult to automate the iterative processes that take place between them, as they develop business collaborations. Contract negotiation is one of such processes, since many interactions may be required until both parties reach an agreement. For these processes there is no standard activity sequence, hence the difficulty of workflow systems in modeling this behaviour. However, in terms of concurrency, it all comes down to the set of messages that organizations typically exchange with each other, and this is a well-known set of interactions that can be easily specified.

Let us assume that, while negotiating a contract, organization *A* prepares and sends a contract proposal to organization *B*. *B* considers the proposal and decides whether it should be accepted or not. If *B* does not accept it, then *A* will revise the proposal and send a new one. This behaviour will repeat itself until *A* and *B* reach an agreement. Figure 2 shows the processes running at both ends, as well as the interactions that take place between them. Basically, there is a cycle running at each end: *A* sends and revises proposals until it gets an affirmative response from *B*; on its turn, *B* receives, analyzes and replies to each proposal until an agreement is reached.

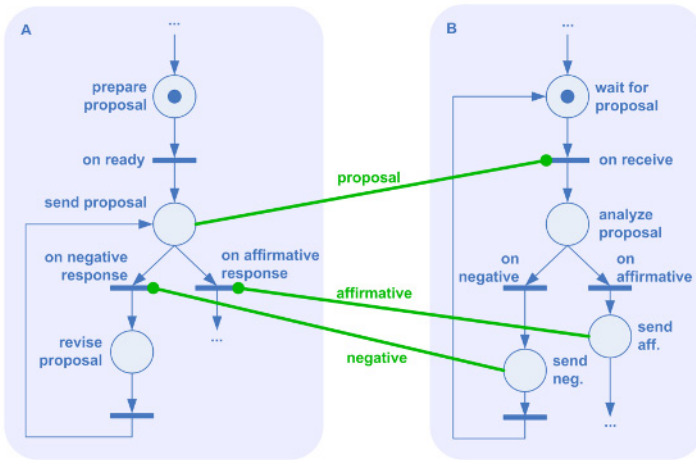


Fig. 2. Sequencing and concurrency in contract negotiation

In terms of concurrency, the above interactions could be described in π -calculus as:

$$A(link) = \overline{link} \langle proposal \rangle . (link(yes) + link(no) . A(link)) \tag{3}$$

$$B(link) = link(proposal) . (\overline{link} \langle yes \rangle + \overline{link} \langle no \rangle . B(link)) \tag{4}$$

4 Using Mobility to Model Partner Search

A distinctive feature of π -calculus when compared to earlier process algebras is the concept of *mobility*, which is intended to allow π -calculus to model processes with changing structure [27]. Basically, mobility involves the ability to send links through links. This mechanism turns out to be extremely useful to model partner search, since it provides a way to describe how organizations find information about other organizations at run-time.

Let us assume that, in an electronic market, participants will have access to a partner search service. This service may be implemented as a centralized repository (such as a UDDI registry) or as a fully decentralized service as in a peer-to-peer e-marketplace [28]. Regardless of how the search service is physically implemented, it can be represented as a provider of links to market participants. The search service receives requests, matches them against the products/services available in the market, and replies with one or more possible candidates.

Figure 3 illustrates the interactions between an organization A and the partner search service. The service is invoked during the execution of a local process at A , which aims at finding a suitable partner for a given business need. The search service replies with a (presumably non-empty) set of matching results, from which A will select an interesting candidate, which will be referred to as organization B . Then A interacts directly with B in order to obtain further information about that candidate and its product offers.

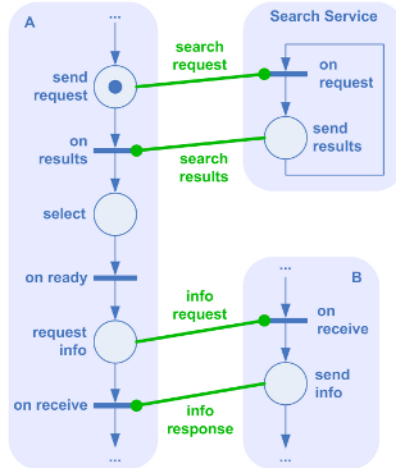


Fig. 3. Modeling partner search with mobility (simplified)

In order to formalize this behaviour using π -calculus, we will assume that the following services are available in the market [29]:

- *Trading Partner Search Service (TPSS)* - the partner search service just described.

- *Trading Partner Information Service (TPIS)* - a service link that allows market participants to exchange institutional or product information in order to support partner selection. In general, TPIS will be used to retrieve additional information about the possible candidates provided by TPSS.
- *Trading Partner Agreement Service (TPAS)* - a service link that allows market participants to negotiate contracts, as described in the previous section. In general, TPAS will be used when a candidate partner has already been selected. The communication channel required by TPAS is obtained via TPIS.
- *Trading Partner Execution Service (TPES)* - a one-to-one service link that allows market participants to perform the interactions established in a previously agreed contract. The communication channel required by TPES is obtained via TPAS.

The interactions shown in figure 3 can then be expressed using the following expressions, where the subscript is used to denote sub-processes belonging to the same entity:

$$PSS(tpss) = tpss(request).\overline{tpss} < result > .PSS(tpss) \quad (5)$$

$$A_{SEARCH}(tpss) = \overline{tpss} < req > .tpss(res).A_{SELECT}(res) \quad (6)$$

$$A_{SELECT}(candidates) = A_{INFO}(selection) \quad (7)$$

$$A_{INFO}(tpis) = \overline{tpis} < tpis_request > .tpis(tpis_response) \quad (8)$$

$$B_{INFO}(tpis) = tpis(tpis_request).\overline{tpis} < tpis_response > \quad (9)$$

5 Case Study: A Semiconductor Supply Chain

In this section, we will illustrate how the proposed approach can be used to model a real-world business scenario. This scenario is basically equivalent to that presented in [30], and it involves three companies from the semiconductor industry. To avoid working with real names, we will refer to these companies simply as *A*, *B* and *C*. Organization *A* is a manufacturer of electronic subsystems for the automotive industry. Most of these components require application-specific integrated circuits (ASICs), which *A* orders from *B*. In order to produce these and other customized integrated circuits, company *B* needs silicon wafers, which are supplied by *C*. Silicon wafers are standard products, which *C* can supply from its own stock, while driving production for stock replenishment.

For the purpose of our scenario, we will assume that none of these companies know each other *a priori*, so that they will have to search for and develop business collaborations with one another. First, company *A* will search for a supplier of ASICs. Having identified a list of potential suppliers, *A* will engage in conversations in order to select the best supplier, which eventually will be company *B*. Then *A* and *B* will sign a contract which specifies how the purchase will take place, from initial ordering to final payment. Company *B* will produce the ASICs and send them to *A*, so that *A* can proceed with its own manufacturing

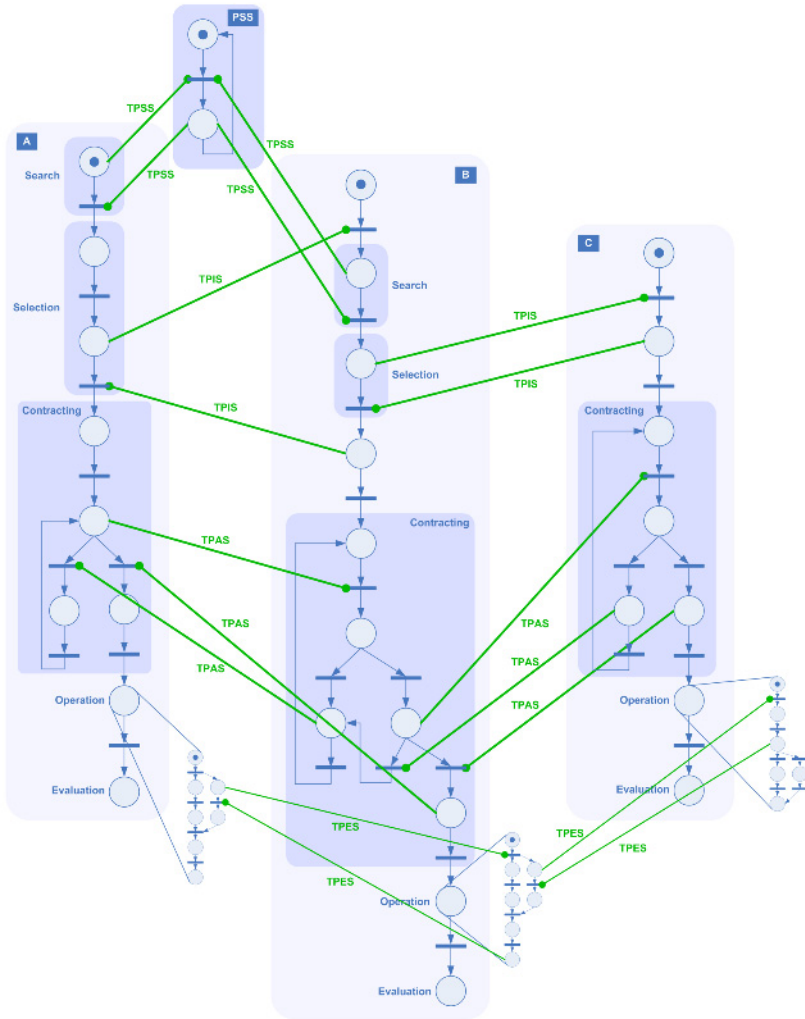


Fig. 4. Concurrent processes in a semiconductor supply chain (simplified)

process. Finally, *A* will measure the performance of the supplier so that this information can be taken into account in future partner selections.

While *A* develops its collaboration with *B*, *B* develops a collaboration with *C* according to a similar procedure. The timing of these events is such that when *A* searches for a supplier of ASICs, *B* searches for a supplier of wafers. And when *A* signs a contract with *B*, *B* signs a contract with *C* simultaneously. In this case, *B* plays both the service-requester and service-provider roles at the same time, on one side towards *C* and on the other side towards *A*, respectively.

During the operation phase, company *A* initiates production of the electronic system and at the same time sends the order for the ASICs to company *B*. Then

B initiates its chip-manufacturing process, while at the same time ordering the silicon wafers. Supplier *C* provides the wafers immediately and replenishes its own stock if needed. The wafers arrive in the middle stages of *B*'s manufacturing process which, when completed, will provide the ASICs to company *A*.

6 Conclusion

Cross-organizational workflow management has been a long way from the run-time problem of supporting distributed workflow execution, up to the build-time problem of decoupling models of public collaborative workflows from those of private, local business processes. Still, it is clear that inter-organizational processes require solutions other than just mechanisms to enforce activity sequencing. The concurrent nature of inter-organizational processes calls for a solution that supports the interconnection of business processes running under the control of different organizations.

In this context, π -calculus becomes a valuable tool to model concurrency and synchronization in inter-organizational processes, just like Petri nets are useful to model activity sequencing in local business processes. In this paper, we have described mainly in a graphical way how these two formalisms can be combined in order to model inter-organizational systems. In the future, we will study the combination of these notations on formal grounds in order to determine the kind of properties that can be formally verified.

References

1. D. Hollingsworth, "The Workflow Reference Model", Document Number TC00-1003, WfMC, 1995
2. R. Anzböck, S. Dustdar, "Interorganizational Workflow in the Medical Imaging Domain", ICEIS'03, Angers, France, April 23-26, pp.292-300, Kluwer, 2003
3. W. van der Aalst, "Process-oriented architectures for electronic commerce and interorganizational workflow", Information Systems, 24(9), December 1999
4. A. Geppert, M. Kradolfer, D. Tombros, "Federating heterogeneous workflow systems", Technical Report 05, University of Zurich, 1998
5. F. Lindert, W. Deiters, "Modelling Inter-Organizational Processes with Process Model Fragments", Informatik'99, Paderborn, Germany, October 1999
6. M. Reichert, T. Bauer, P. Dadam, "Enterprise-Wide and Cross-Enterprise Workflow Management: Challenges and Research Issues for Adaptive Workflows", Informatik'99, Paderborn, Germany, October 1999
7. M. Blake, "An Agent-Based Cross-Organizational Workflow Architecture in Support of Web Services", 11th WETICE, IEEE Comp. Soc. Press, June 2002
8. M. Kwak, D. Han, J. Shim, "A Framework Supporting Dynamic Workflow Interoperation and Enterprise Application Integration", 35th HICSS, IEEE, 2002
9. C. Stricker, S. Riboni, M. Kradolfer, J. Taylor, "Market-based Workflow Management for Supply Chains of Services", 33rd HICSS, IEEE, 2000
10. S.-B. Yan, F.-J. Wang, "CA-PLAN, an Inter-Organizational Workflow Model", FTDCS'04, IEEE Computer Society, 2004

11. P. Grefen, K. Aberer, Y. Hoffner, H. Ludwig, "CrossFlow: cross-organizational workflow management in dynamic virtual enterprises", *Comp. Sys. Sci. and Eng.*, 15(5):277-290, 2000
12. W. van den Heuvel, H. Weigand, "Cross-Organizational Workflow Integration with Contracts", *Business Obj. Comp. Workshop (OOPSLA2000)*, Springer, 2001
13. E. Kafeza, D. Chiu, I. Kafeza, "View-based Contracts in an e-Service Cross-Organizational Workflow Environment", *VLDB-TES, Rome, Italy*, 2001
14. W. van der Aalst, M. Weske, "The P2P approach to Interorganizational Workflows", *CAiSE'01, LNCS 2068*, pp.140-156, Springer, 2001
15. W. van der Aalst, T. Basten, "Inheritance of Workflows: An Approach to Tackling Problems Related to Change", *Theo. Comp. Sci.*, 270(1-2):125-203, 2002
16. I. Chebbi, S. Dustdar, S. Tata, "The view-based approach to dynamic inter-organizational workflow cooperation", *Data and Knowledge Eng.*, 56(2), 2006
17. D. Chiu, S. Cheung, S. Till, K. Karlapalem, Q. Li, E. Kafeza, "Workflow View Driven Cross-Organizational Interoperability in a Web Service Environment", *Information Technology and Management*, 5(3-4), July-October 2004
18. E. Cimpian, A. Mocan, "WSMX Process Mediation Based on Choreographies", *1st Intl. Workshop on Web Service Choreography and Orchestration for Business Process Management*, Nancy, France, September 2005
19. H. Smith, P. Fingar, "Workflow is just a Pi process", *BPTrends*, January 2004
20. W. van der Aalst, "Pi Calculus Versus Petri Nets: Let Us Eat Humble Pie Rather Than Further Inflate the Pi Hype", *BPTrends*, 3(5), pp.1-11, May 2005
21. F. Puhlmann, M. Weske, "Using the Pi-Calculus for Formalizing Workflow Patterns", *BPM 2005, LNCS 3649*, pp.153-168, Springer, 2005
22. D. Yang, S. Zhang, "Approach for workflow modeling using pi-calculus", *Journal of Zhejiang University Science*, 4(6):643-650, Nov-Dec 2003
23. J. Padget, J. Bradford, "A pi-calculus Model of a Spanish Fish Market - Preliminary Report", *AMET-98, LNCS 1571*, pp.166-188, 1998
24. D. Ferreira, J. J. Pinto Ferreira, "Developing a Reusable Workflow Engine", *Journal of Systems Architecture*, 50(6):309-324, June 2004
25. W. van der Aalst, "The Application of Petri Nets to Workflow Management", *The Journal of Circuits, Systems and Computers*, 8(1), pp.21-66, 1998
26. D. Ferreira, "Workflow Management Systems Supporting the Engineering of Business Networks", *PhD Thesis, University of Porto*, February 2004
27. R. Milner, J. Parrow, D. Walker, "A calculus of mobile processes, Part I", *Information and Computation*, 100(1):1-40, September 1992
28. D. Ferreira, J. J. Pinto Ferreira, "Essential Services for P2P e-Marketplaces", *CARS&FOF 2002, Porto, Portugal*, July 3-5, 2002
29. D. Ferreira, J. J. Pinto Ferreira, "Building an e-marketplace on a peer-to-peer infrastructure", *Intl. J. of Comp. Int. Manufact.*, 17(3):254-264, April-May 2004
30. D. Ferreira, J. J. Pinto Ferreira, "Towards a workflow-based integration architecture for business networking", *Business Process Management Journal*, 11(5):517-531, 2005

From Inter-organizational to Inter-departmental Collaboration – Using Multiple Process Levels

Daniel Simonovich

Reutlingen University, Alteburgstrasse 150, D-72762 Reutlingen
daniel.simonovich@reutlingen-university.de

Abstract. Business process modeling has gained fundamental importance for describing the collaboration in and between enterprises. However, the lack of standardization and range of levels of detail used in corporate practice pose practical challenges. This is particularly true in transformations like outsourcing or mergers and acquisitions, where a high degree of flexibility is required from both organization structures and information and communication technology (ICT). Rather than suggesting a process modeling standard to fit all purposes, this paper presents three complementary and reality-checked levels of process design tailored at different audiences and serving corresponding modeling purposes in transformation projects.

Keywords: Business process management, conceptual modeling, enterprise collaboration, organizational transformation.

1 Introduction

Ever since the 1990s, it became clear to the business community that not only the right choice of markets and products mattered, but that the ability to execute superior internal activities counted as well. Popularized by best-selling books [1], this realization was captured by the notion of business processes, a series of steps aimed at producing a specific result [2], [3], [4]. Owing to the need of automating business process flows in and between organizations through information and communication technology (ICT), IS professionals developed formal methods. Largely inspired by Petri net theory and owing to the sustained importance of business processes, the academic IS community not until recently introduced dedicated annual workshops to promote intensified formal research [5]. Most of the formal specification output, however, reflects detailed and rather systems oriented procedures at the cost of neglecting higher level process descriptions as used throughout the business and consulting communities. Without departing from the conceptual modeling tradition underlying contemporary research advances, a set of three process description layers, popularly used in practice, is presented and expressed in mathematical terms, thereby inviting the formal extension of detailed process nomenclature with higher level instantiations used for managerial analyses. The process levels presented are referred to different collaboration scenarios as well as to major transformational initiatives.

2 Three Levels to Cover Real World Business Process Practice

Business processes have changed the way business and ICT professionals interpret organizational problems. However, with little advice on how to execute process analysis in early publications, the process movement led to an array of mapping versions. In the following, three archetypical de-facto standards are reviewed.

2.1 The Practical Need for Multiple Levels

The definition of a business process as a set activities aimed at producing a specific result [2] is as vast as ranging from the entirety of steps to develop a new product to a single mouse click completing an internet order. While the crafting of higher level processes requires rather superficial and easy-to-glimpse mapping, systems oriented process descriptions strive to eliminate uncertainty by specifying in-depth flow logic and information retrieval context. Consequently, the audience for process descriptions ranges from a firm's senior management to systems development professionals. It is therefore evident that different addressees require dissimilar modeling constructs. It should be mentioned that detailed tool sets such as the widespread ARIS instrument usually offer aggregate views [6] and are sporadically extended by value creation logic [7]. However, these top-level representations are far from representing the kind of maps used in managerial reality. The next subsection selects a set of process map traditions covering the extensive range of process design purposes and audiences.

2.2 A Reality Checked Set of Process Model Descriptions in Use

Highest level process analysis and design is little different, if not coinciding, with the celebrated value chain concept [8]. Being the earlier framework, the value chain is different in purpose from the business process paradigm in that it strives to give a complete picture of a firm's activities, suitable for analyzing the strengths and weaknesses of a firm [9]. Business processes, on the other hand, usually deal with a specific set of activities. However, a growing strand of research suggests that the two approaches are not in competition with one another but rather complementary [10], sometimes using the value chain as the top level reference for further breakdown into business processes. Both paradigms have produced similar high level maps restricted to linear sequence logic, as exemplified by the E-Business planning process of Terra Lycos [11] (see Fig. 1). This is the first process compound suggested in a set of three representative levels. At the opposite end of the spectrum, event-driven process chains (EPCs) are in popular use as well as subject of ongoing research [12]. Usually far too detailed for management, EPCs use a rich set of proprietary symbols to navigate from state to state through characterized activities. In between simple high level maps and EPCs lives a popular mapping style somewhat deserted from academic treatment. This sandwich-level representation bears no standard notion and is henceforth referred to as "organizational flow chart". Such maps emphasize where activities actually take place, thus being ideal for illustrating inter-departmental process flow logic. Fig. 2 shows a piece of real world process modeling taken from a case study of the service provider paybox.net [13]. As suggested in Fig. 3, all three business process

design levels are in heavy real-world use, propagated by consulting professionals in the areas of strategy, process design, and ICT [14]. In the following, these established notations will be linked to present-day conceptual modeling research.

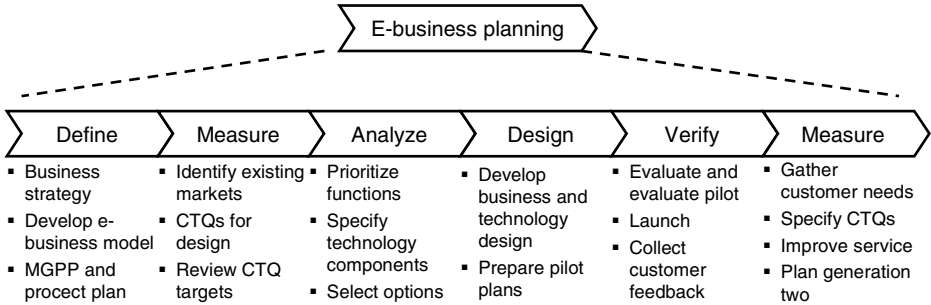


Fig. 1. High style map of planning process at Terra Lycos

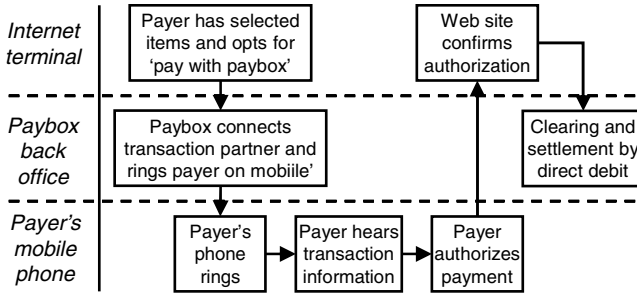


Fig. 2. Organizational flow chart for the internet-to-paybox payment process of paybox.net

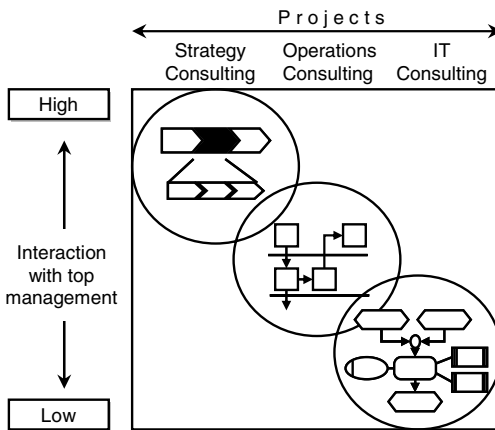


Fig. 3. Types of process maps prototypical for the range of consulting engagements

3 Translating Real World Practice into Formal Conceptual Models

Of the three process levels introduced, only the lowest has become a distinct subject of intensive formal treatment in research [12]. Nevertheless, the relevance of the higher levels and their power to dictate conditions to the third level, suggests their formal treatment, too. Subsequently, a conceptual model is developed to integrate all levels in the formal process modeling tradition. With its general-purpose outlook, graph theory is preferred over more purpose-specific approaches such as Petri nets.

3.1 Building the Conceptual Model Set

Each level is modeled in turn before integration into an overall framework. The highest level presents a mere sequence of steps which may be further broken down into subsequences, whereas, a breakdown into activities may not necessarily be sequenced. This translates into a connected tree structure with optional lists of sub steps sharing the same parent. Let (V, E_V) be a digraph satisfying a tree property:

$$\exists v \in V : g^-(v) = 0 \wedge (\forall u \in V : u \neq v \Rightarrow g^-(u) = 1), \quad (1)$$

whereas g^- denotes the in-degree of a node, and the connectedness of the graph holds when for all $(u, v) \in E_V$:

$$\exists n > 0 : \exists u = u_1, u_2, \dots, u_n, = v : \forall i < n : (u_i, u_{i+1}) \in E_V \cup E_V^{-1}. \quad (2)$$

The fact that sequential relationships may exist only between process steps sharing the same activity parent is captured by a path $\langle v_1, v_2, \dots, v_n \rangle$ satisfying

$$\exists u \in V : \forall i : 1 \leq i \leq n : (u, v_i) \in E_V. \quad (3)$$

Turning to the organizational flow chart, the formal logic can be captured by a digraph (P, E_P) , providing each process step $p \in P$ with a department of the departmental set D .

$$f_D : P \rightarrow D. \quad (4)$$

Event-driven process chains (EPCs) are used to exemplify modeling at the lowest level. Let S be the set of states, A be the Set of activities and C be the set of connectors. Then an EPC can be represented by a digraph (M, E_M) with

$$M = S \cup A \cup C \quad \wedge \quad E_M \subseteq M \times M, \quad (5)$$

whereas states, activities and connectors alternate and no two activities are joined by a connector:

$$S^2 \not\subseteq E_M \wedge A^2 \not\subseteq E_M \wedge C^2 \not\subseteq E_M \wedge \forall a_1, a_2 \in A : (\neg \exists c \in C : (a_1, c) \in E_M \wedge (a_2, c) \in E_M). \quad (6)$$

The connectors denominate the logical symbols of “and”, “or”, and “exclusive or”. This semantic can be maintained by the function

$$f_C : C \rightarrow \{\wedge, \vee, \otimes\}. \quad (7)$$

Organizational units and informational references are captured by the functions

$$f_U : A \rightarrow U, \quad f_I : A \rightarrow I. \quad (8)$$

It should be mentioned that different versions of EPCs exist leading to different formalization complexity [7]. Note that the descriptions so far are less cumbersome than existing EPC formalizations [12], even at the inclusion of higher level process design nomenclature.

3.2 Addressing the Need for Consistency

All the sets defined above including the descriptions (1)-(8) create an overall conceptual model capturing all three process design levels simultaneously. However, no links have so far been created between these three models. It seems reasonable to demand that a higher level process activity represents a number of lower level activities, furthermore ensuring that no higher level activity is not fleshed out in lower level details (surjectivity). This logic can be satisfied using two functions:

$$f_{AP} : A \rightarrow P, \quad f_{PV} : P \rightarrow V, \quad (9)$$

which satisfy surjectivity:

$$\forall p \in P \exists a \in A : f_{AP}(a) = p \wedge \forall v \in V \exists p \in P : f_{PV}(p) = v. \quad (10)$$

Similar consistency conditions may be expressed to address compatibility of organizational departments and units or logical flow relationships between the levels.

After the possibility of fusing together a multi-level process model from existing real world practice has been demonstrated using formal logic, we turn to the practical settings that require these different process levels.

4 Using Multiple Levels for Different Basic Collaboration Scenarios

The need for organizational agility on one hand and the increasing sophistication of information systems and ICT infrastructures on the other prompts enterprises to understand and master a broad range of collaborative situations and constellations. The next two subsections deal with both basic and transformation scenarios. Using argumentation, it is put forward when the three business process design techniques are likely to be in use and how they might be deployed in major change initiatives.

4.1 Basic Enterprise Collaboration Scenarios

A useful reference for distinguishing collaborative constellations from an organization perspective is Porter's value chain [8]. Besides intradepartmental collaboration, interdepartmental relationships exist either through the coordination of primary activities or support provided by secondary to the primary activities. Beyond the boundaries of an organization, different relationships characterize inter-organizational

collaboration. These include outsourcing of non core activities [15], supply chain interaction through a chain of inbound and outbound logistics relationships [16], and finally, corporate customer relationships [17] where the areas of marketing and sales interact with a customer’s purchasing and other departments. The various scenarios are depicted in Fig. 4.

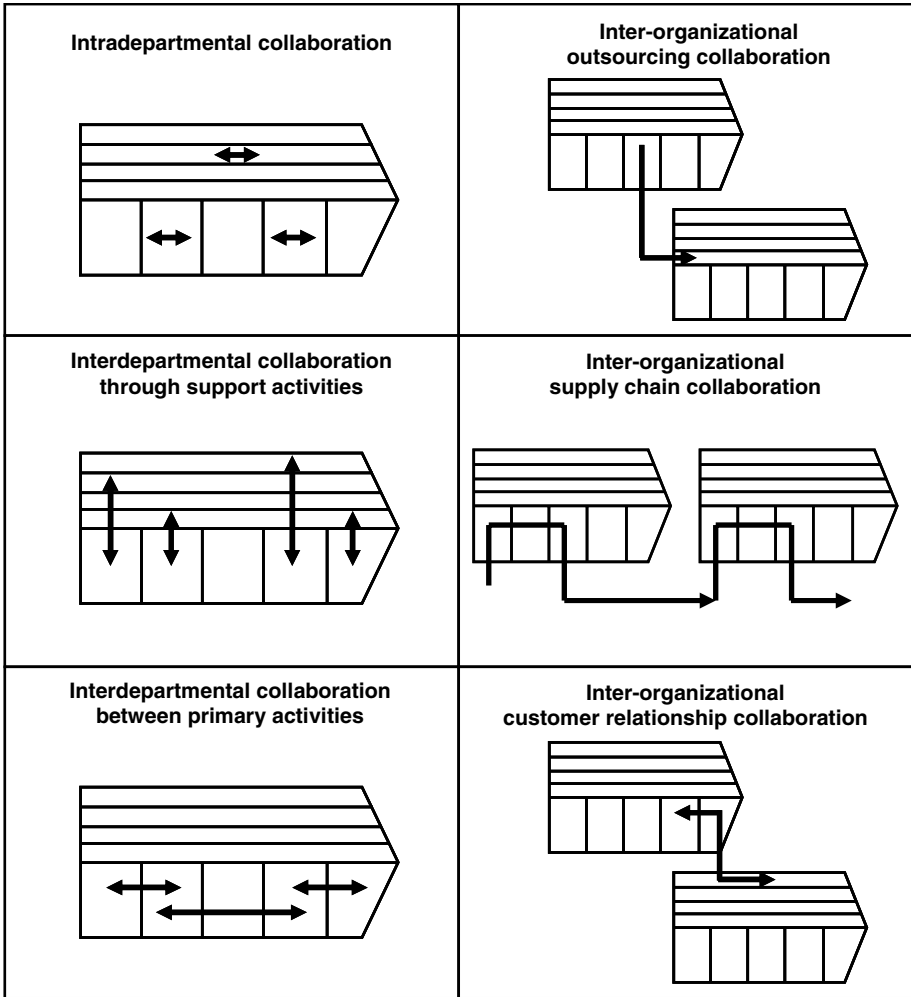


Fig. 4. Basic collaboration scenarios

Fig. 5 refers these basic collaboration scenarios to the three process modeling levels introduced. The value-chain-style first process level is suitable for understanding the rationale for any inter-organizational relationship including the division of value creation and costs among different corporate actors. However, to truly understand how inter-organizational collaboration flows between the parts of

two enterprises, an organizational flow chart can be highly effective, whereas the viability of event-driven process chain (EPCs) is limited to the very specifics of inter-organizational activities. Turning to process activities within an enterprise, the first process level loses any attractiveness beyond scope setting while the organizational flow chart referred to as “level two” displays its true strengths in serving as the ultimate interdepartmental collaboration map. Clearly, EPCs tend to become hardly manageable when utilized beyond a specific application area.

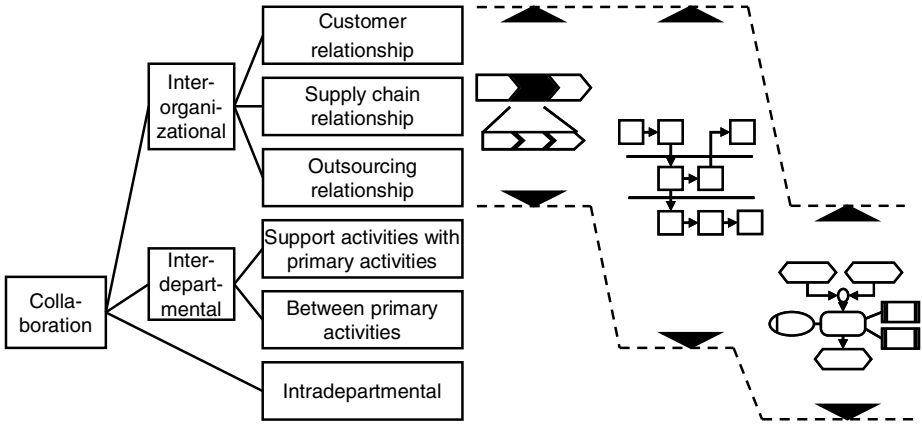


Fig. 5. Referring the process design levels to a taxonomy of enterprise collaboration

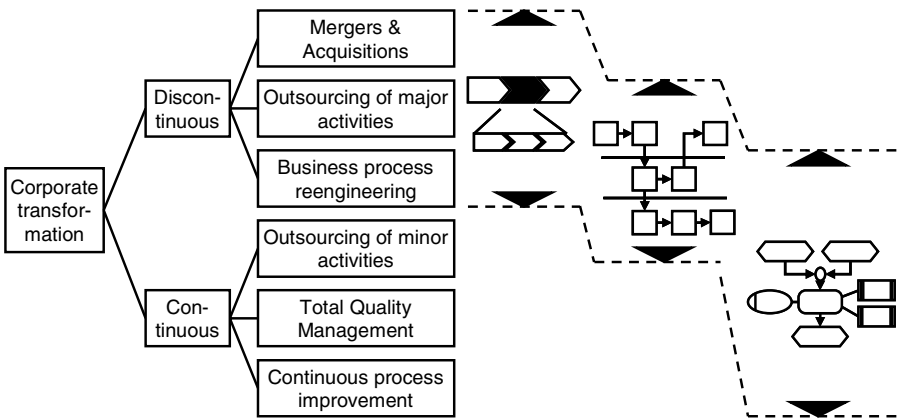


Fig. 6. Referring the process design levels to a taxonomy of enterprise transformation

4.2 Transformational Enterprise Scenarios

Corporate transformations can be characterized as either continuous (first order change) or discontinuous (second order change) [18]. Continuous change initiatives

such as continuous improvement [19] or total quality management [20] rely on subtle process improvements. Generally, detailed process description techniques such as EPCs are well suited to capture the sometimes microscopic facets of gradual change. In reengineering initiatives [1], these formal descriptions are useful to specify the final state of a creative process rather than supporting the dynamic testing out of radical change creativity midway. Second-level organization charts are useful when new ways of collaboration between different organizational units require specification. The highest level process map may be suited in any discontinuous change initiative, being a constant point of reference in an otherwise context of uncertainty.

5 Conclusions

Although the business process paradigm has existed for well over a decade, research efforts re-accelerated over the last three years to exploit the potential of formal business process modeling [5]. Unfortunately, much of the mathematical work inspired by Petri net theory is remote from the real world use of process maps to solve critical organizational problems. Using event-driven process chains (EPCs), arguably the most prominent example of detailed conceptual modeling in professional use, two higher level process map types were complemented to yield realistic sets of process maps to arrive at three levels of organizational detail. Referring to a number of collaborative scenarios, it was demonstrated how these different process maps might be balanced. While the first level is suitable for showing how an organization's mission breaks down into critical value chain steps, the intermediate "sandwich" level covers a broad range of collaborative and transformational settings. Lowest level process modeling, as exemplified by EPC notation, is unwieldy for transformational circumstances but fundamental to articulate systems implementation clarity. It should not remain unmentioned that commercially available EPC tools such as ARIS [6] offer summarizing abstractions, similar in purpose to the higher level process maps presented in this paper. However, these proprietary aids simply do not look like the higher level process maps used by business professionals or strategy consultants. The formal conceptual treatment of the process levels above EPCs suggests that there is little reason to exclude practical and easy-to-understand modeling constructs, as used in the real world, from formal research work. Quite on the contrary, disregarding well-established process design traditions used by business professionals bears the risk of increasing the understanding gap between the business and ICT communities.

Opportunity for further research: First, tool construction is a natural follow-up of the unifying formal framework, either from scratch or by extension of existing single-style instruments. On the empirical edge, evidence may be sought for finding the ideal process level mix by industry, business function, or stakeholder group. Finally, further conceptual research may probe into the integration of strategic management instruments through their relationship with higher process levels. Any such work would aim for an improved business-ICT alignment, a prime concern of CIOs [21].

References

1. Hammer, M., Champy, J.: Reengineering the corporation. Harper, New York (1993) 35-39
2. Davenport, T.H.: Process Innovation: Reengineering Work through Information Technology. Harvard Business School Press, Boston (1993) 5
3. Hammer, M.: Don't Automate, Obliterate, Harvard Business Review, July-August (1990) 104-112
4. Morris, D., Brandon, J.: Reengineering your Business, McGraw Hill, New York (1993)
5. Aalst, W.M.P., Hofstede, A., Weske, M. (eds.): International Conference on Business Process Modeling BPM 2003, Proceedings. Springer, Berlin Heidelberg New York (2003)
6. Scheer, A.-W.: Business Process Engineering: Reference Models for Industrial Enterprises. Springer, Berlin Heidelberg New York (1999)
7. Neiger, D., Leonid C.: Goal-Oriented Business Process Modeling with EPCs and Value-Focused Thinking. Second International Conference Business Process Modeling BPM 2004, Springer, Berlin Heidelberg New York (2004) 98-115
8. Porter, M.E.: Competitive Advantage – Creating and Sustaining Superior Performance. Free Press, New York (1985) 33-61
9. Hax, A.C., Majluf N.S.: The strategy concept and process, 2nd edition. Prentice Hall, Upper Saddle River (1996) 118-127
10. Jelassi, T., Enders, A.: Strategies for E-Business – Creating Value through Electronic and Mobile Commerce. Pearson (2005) 114-116
11. Terra Lycos, Internal Company Presentation, August (2001) 12
12. Aalst, W.M.P.: Formalization and Verification of Event-Driven Process Chains. Information and Software Technology, Vol. 41, No 10 (1999) 639-650
13. Jelassi, T., Enders, A.: Strategies for E-Business – Creating Value through Electronic and Mobile Commerce. Pearson (2005), 570-574
14. Téoul, J.: We are all in services now, INSEAD working paper, Fontainebleau (1997)
15. Lacity, M., Willcocks, L., Feeny, D.: IT Outsourcing: Maximize Flexibility and Control. Harvard Business Review, May-June (1995) 84-93
16. Vollmann, T.E., Vordon, D., Raabe, H.: Supply Chain Management. In: Dickson, D., Bickerstaffe, G. (eds.): Mastering Management, Pitman, London (1997) 316-322
17. Romano, M.C., Fjermestad, J.: Electronic Commerce Customer Relationship Management: A Research Agenda. Information Technology and Management, Vol. 4 (2003) 233-258
18. Hersey, P., Blanchard, K., Johnson, D.E.: Management of Organizational Behavior: Leading Human Resources, 8th edition, Prentice Hall, Upper Saddle River (2000), 387-389
19. Cordon, C.: Ways to Improve the Company. In: Dickson, D., Bickerstaffe, G. (eds.): Mastering Management. Pitman, London (1997), 307-311
20. Krajewski, L.J., Ritzman, L.P.: Operations Management: Strategy and Analysis, 4th edition. Addison-Wesley, Reading (1996) 139-178
21. Luftman, J., Kempaiah, R., Nash, E.: Key Issues for IT Executives 2005. Management Information Systems Quarterly - Executive, Vol.5, No.2 (2006)

An Ontology-Based Scheme Enabling the Modeling of Cooperation in Business Processes

Manuel Noguera, M. Visitación Hurtado, and José L. Garrido

Dpt. Lenguajes y Sistemas Informáticos, University of Granada,
E.T.S.I. Informática, C/ Periodista Daniel Saucedo Aranda,
18015 Granada, Spain
{mnoguera, mhurtado, jgarrido}@ugr.es

Abstract. Nowadays enterprises face more and more ambitious projects which require the cooperative participation of different organizations. The modeling of one organization information is a crucial issue. There are several approaches for analyzing or representing how an organization is structured and operates. Much recent research points to ontologies as an appropriate technology for modeling enterprise systems. This paper presents a proposal for modeling cross-enterprise business processes from the perspective of cooperative systems. The proposal consists of a multi-level design scheme for the construction of cooperative system ontologies, and it is exemplified through a real case study. Benefits related to ontology integration are also presented.

Keywords: Ontologies, Cross-enterprise Business Processes, Conceptual modeling, CSCW.

1 Introduction

Enterprise modeling techniques aim at modeling the behavior and domain entities in order to identify the fundamental business principles of an organization [1, 11, 16]. Modeling both structural and behavioral aspects in an integral manner is a challenging issue, since it is difficult to represent both aspects at the same time, especially when two or more companies cooperate in business processes. In order to support enterprise cooperation and integration it is necessary that shareable representations of enterprise business process are available. This will minimize ambiguity and maximize understanding and precision in communication.

As a result of technological evolution, enterprises must continuously rethink their business designs, and organizations must change their enterprise models. In this context, one key issue is the modeling of intra- and inter-organizational cooperation in order to achieve useful and sharable knowledge representations. The objective is to express formally how the enterprise works in its own terminology, and to enable the enterprise to exchange processes, rules, groups, etc.

The Computer-Supported Cooperative Work (CSCW) discipline [10] can play an important role in enabling cross-enterprise business processes as it studies cooperation across organizations that use new technologies. Although an enterprise model embraces various aspects (marketing, costs, strategy, etc.) [11], due to the nature of the cooperation processes, these processes should be integrated into an

enterprise activity model, which in turn might be directly connected to an organization model [9].

In order to provide enterprise models with a common, formal vocabulary and semantics, much recent research points to ontologies as an appropriate technology for this purpose [5, 12]. Furthermore, one added value of formalizing ontologies with standard languages (e. g. OWL-Web Ontology Language [23]) is that they support machine-processable descriptions of the models and reasoning based on Description Logics [4]. In this respect, they may provide some kind of guidance during system specification in order to help prevent mistakes being made. Likewise, the conformity with a predefined conceptual schema of the ontological description of a system might be automatically checked.

This research work focuses on the modeling cross-enterprise business processes under the perspective of cooperative systems. The aim is to represent, in an integrated manner, the structure and dynamics of groups belonging to different organizations that participate in cooperative tasks. In order to accomplish this, we leverage a domain ontology from which to start defining application ontologies [12]. An application ontology describes *“concepts depending on a particular domain [...] These concepts correspond to roles played by domain entities while performing a certain activity”* [12]. In our case the domain ontology would describe the terms involved in the description of a cooperative system (such as “role”, “task”, “group”, “activity”, “organization”, etc.).

In a cooperative system, it is usual that one actor may play several roles and change from one to another at any time. In this context, certain changes in the application ontology (e. g. an actor that leaves a role to start playing another one) are more likely to occur than others (e. g. adding a new role or suppressing an outmoded task, etc.), and even more so in the domain ontology (e.g. change the definition of a role as “a set of tasks” or that an actor is part of a group). An appropriate multi-level design of the underlying ontologies has been taken into account in order to allow certain changes to be made at the corresponding level without interfering with the others. Even at the application ontology level, yet more different levels could be distinguished depending on the intrinsic dynamism of the entities they describe and their generality/specificity.

As a leading example, we will consider the process of granting a mortgage in a branch office. Different actors belonging to different organizations are involved, but they are all part of the same group in charge of the granting process. In terms of the organizations, there are a “Branch”, a “Valuation Office” and a “Notary Office”.

The remainder of this paper is organized as follows. Section 2 gives an overview of ontology design issues. Section 3 presents the conceptual framework used as starting point for our proposal. Section 4 introduces our ontology-based design scheme for modeling cooperation in business processes. Finally, main conclusions are presented in Section 5.

2 Previous Considerations in Ontology Design

The objective of this work is to provide the foundations for describing cooperation processes between enterprises using ontologies. In this section we summarize some of

the problems which normally appear in designing ontologies and which we will try to overcome with our proposal.

Despite the amount of methodologies, tools and languages proposed in order to design and build ontologies [6, 19], the construction of easy-to-integrate, reusable and “validatable” ontologies continues to be an unsolved problem [14, 18]. Many of the current ontology-based descriptions of a system or definitions of concepts may be classified as “ad-hoc ontologies” (i.e. they have been elaborated for a particular environment with its particular context, thereby hindering their subsequent reuse). A machine does not know when two concepts (or relations) are equivalent (or when the set of objects of a certain class are disjoint with the others) unless this has been explicitly defined. This requires an additional effort for the creation of a new derived ontology to check whether each of its terms appropriately fits the new context. In our case, we focus on the elements needed to model the group and the organization of the groupwork.

At the same time, these systems and organizations change. The evolution of the ontology describing a system is another important issue. Most of the changes usually affect local parts of the system or organization [15]. Ontology design should allow the modifications to be managed without affecting those elements that are unrelated to these changes and without making it necessary to put them “out of service”.

3 A Conceptual Framework for Cooperative Systems

There are several approaches for analyzing or representing how an organization is structured and operates [1]. Nevertheless, a variety of terms are used fairly interchangeably to describe a cooperative environment and organization functions. People involved in the design of complex systems (as cooperative systems are) do not often agree on the terms used to talk about the entities that may appear in the organization. Furthermore, even when the same terms are used, the meanings associated with them may differ, i.e. the semantics.

A conceptual framework is therefore needed to exchange business process specifications between business people using a common vocabulary. The resulting specification must be sufficiently clear, unambiguous and readily translatable into other representations. The final outcome must be a description of the problem domain entities (in our case this would be the cooperative systems domain).

The description of the system must be general enough to allow a wide variety of cooperative systems to be modeled so that it can be reused. In order to avoid suffering from the same drawbacks as ad-hoc ontologies, we have taken the conceptual framework devised by AMENITIES [9], a methodology that studies cooperative systems. Then, we have formalized this conceptual framework as a domain ontology [12]. This domain ontology is the starting point to define subsequent application ontologies. Figure 1 shows the conceptual framework for cooperative systems domain, represented using a UML class diagram [24].

According to this conceptual framework, an *action* is an atomic unit of work. Its event-driven execution may require/modify/generate explicit information. A *subactivity* is a set of related subactivities and/or actions. A *task* is a set of subactivities intended to achieve certain goals. A *role* is a designator for a set of

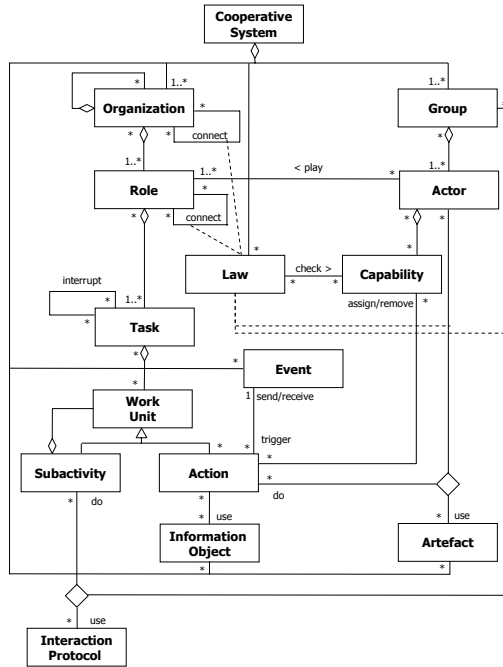


Fig. 1. AMENITIES conceptual framework for cooperative systems

related tasks to be carried out. An *actor* is a user, program, or entity with certain acquired capabilities (skills, category, and so forth) that can play a role in executing (using artifacts) or responsible for actions. A *group* performs certain subactivities depending on interaction protocols. A *cooperative task* is one that must be carried out by more than one actor, playing either the same or different roles. A *group* is a set of actors playing roles and organized around one or more cooperative tasks. A group may consist of related subgroups. A *law* is a limitation or constraint imposed by the system that allows it to dynamically adjust the set of possible behaviors. An *organization* consists of a set of related roles. Finally, a *cooperative system* consists of groups, organizations, laws, events, and artifacts. These concepts and their relations are also applicable to organization and activity models within enterprise models [8], assuming that enterprises are usually engaged in cooperation processes.

4 A Three-Tier Proposal for Designing Cooperative System Ontologies

Based on the philosophy of the approach proposed by Guarino [12], we have devised a three-tier scheme for the design of ontologies for cooperative systems. At the highest level, we would place a document that defines the concepts of the AMENITIES conceptual framework (Figure 1). This would correspond to the domain ontology. At the second level, other documents would appear defining elements

which are specific to each particular system (e.g. “amenities:Organization” and “amenities:Role” labeled boxes), but with the sufficient abstraction level so that they can be employed in similar or related systems (e.g. in our branch cooperative environment the “Valuer” and “Notary” roles). At the third level, yet more specific entities of the system we are dealing with would be declared (e. g. “Anna Riemann”, “Donald Johnson” system actors). Figure 2 shows the outline of the design proposed.

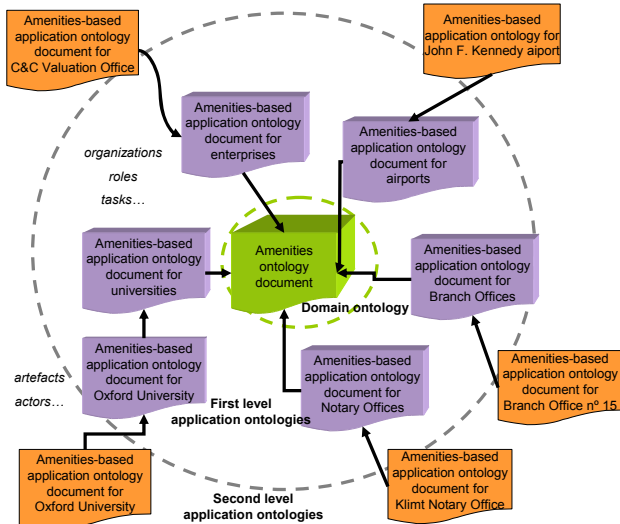


Fig. 2. Three-tier ontology design for cooperative systems

The entities to be declared in either application ontology level will depend on their particular features and not on the type of the entity they refer to. Therefore, structural and behavioral aspects may be defined in either level. For example, in the branch cooperative system, it may be interesting to define a law governing the connection between the “Head of Risk” and the “Bank Manager” when an actor playing the second is absent at a more general level, so that it can be reused in other cooperative systems for other branch offices. On the other hand, a law that rules when “Julie Knowles” (a “Head of Risk”) may play the role “Bank Manager” is only meaningful in the specific system of the “Branch” that Julie works for. The roles “Head of Risk” and “Bank Manager” are general enough to be defined at the first level application ontology. The statement that defines “Julie Knowles” as the “Head of Risk” of the “Branch no. 15” should be declared at the second level application ontology.

The domain ontology described in section 3 may be used to prevent relations being defined between domain elements that do not comply with AMENITIES framework (e.g. defining a group as consisting of a set of tasks, instead of as a set of actors). Thereby some guidance is provided for reducing the likelihood of errors [7] during ontology construction. On the other hand, by using a common conceptual framework, the integration of different cooperative system ontologies would be easier since all would fit the cooperative system pattern proposed in the domain ontology.

4.1 Concepts and Relations in Cooperative Business Processes

If we have a look at the entities that appear in the conceptual framework shown in Figure 1, it can be seen that some of them are not usually modeled as first-class objects. The work presented in [22] adopts a similar approach to ours in the sense that events are modeled as entities (just like any other entities concerning actors, roles, tasks, etc.) which results in a behavioral modeling enhancement.

```

<!-- Domain ontology for AMENITIES conceptual framework -->
<owl:Class rdf:about="#CooperativeSystem">
  <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="hasPart"/>
      </owl:onProperty>
      <owl:allValuesFrom><owl:Class>
        <owl:unionOf rdf:parseType="Collection">
          <owl:Class rdf:ID="Artefact"/>
          <owl:Class rdf:about="#Event"/>
          <owl:Class rdf:ID="Group"/>
          <owl:Class rdf:ID="Law"/>
          <owl:Class rdf:ID="InformationObject"/>
          <owl:Class rdf:ID="Organization"/>
        </owl:unionOf>
      </owl:Class></owl:allValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:about="#hasPart"/>
      </owl:onProperty>
      <owl:minCardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
  </owl:Class>...
  <owl:Class rdf:ID="Role">
    <rdfs:subClassOf><owl:Restriction>
      <owl:allValuesFrom><owl:Class rdf:about="#Organization"/>
    </owl:allValuesFrom>
    <owl:onProperty><owl:TransitiveProperty rdf:about="#partOf"/>
    </owl:onProperty>
    <owl:Restriction></rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction><owl:allValuesFrom>
        <owl:Class rdf:about="#Task"/>
      </owl:allValuesFrom><owl:onProperty>
        <owl:ObjectProperty rdf:about="#hasPart"/>
      </owl:onProperty></owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
    <rdfs:subClassOf>
      <owl:Restriction><owl:onProperty>
        <owl:ObjectProperty rdf:about="#hasPart"/></owl:onProperty>
        <owl:minCardinality rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</owl:minCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>...
  </owl:Class>...
  
```

Fig. 3. Excerpt of domain ontology for cooperative systems according to the AMENITIES conceptual framework implemented using Protégé [20]

It should be noted that in addition to modeling events as first class objects, in AMENITIES the same applies to laws and capabilities. These concepts usually appear encoded in software applications or are modeled through constraint languages [7, 22] such as OCL (Object Constraints Language) [24]. By conceptualizing them, we also enable analysts to model how they relate and affect the other entities in the framework (this is something useful to discuss with users). By using an ontology language for its translation and formalization, we enable it to be machine processed. Figure 3 shows an excerpt of the domain ontology document which we have formalized for cooperative systems in OWL.

It must also be mentioned that laws govern most of the relationships which appear in the conceptual framework in Figure 1. For example, the “connect” relation that enables an actor to change the role that he/she is currently playing is rule-governed by a law in the cooperative system. The law itself is part of the cooperative system and will be instantiated in an application ontology depending on its particular context, but at the same time, it is linked to the association it governs. Laws can therefore be considered not only third-party elements of these sorts of relations but also first class objects.

Unlike diagrammatic descriptions, plain text descriptions of cooperative system elements (such as the one shown in Figure 3) may not be suitable for discussing with certain stakeholders during the business process modeling. Since both types of representations are isomorphic and their expressive powers are similar [3], it is desirable that the translations between them can be carried out automatically. Many of the current tools for building ontologies already support the transformation of an ontology description into a graphical representation. Figure 4 shows the graphical representation for the domain ontology description in Figure 3. The mapping between Figures 1 and 4 is self-evident.

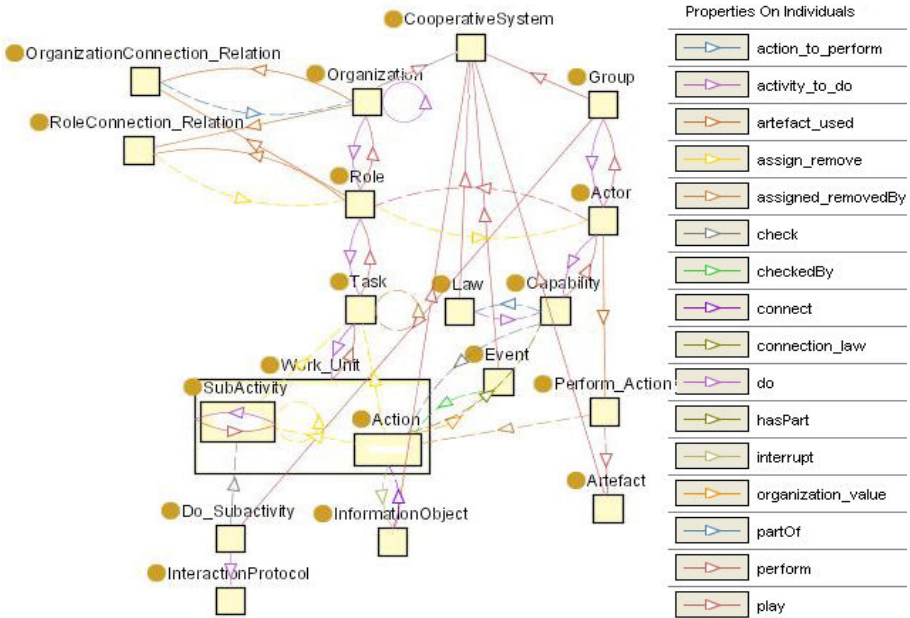


Fig. 4. Transformation of domain ontology description in OWL to a diagrammatic representation. Figure generated using the Jambalaya plug-in [17] for Protégé [20].

4.2 Change Management

It is clear that change management is not a trivial matter. One main drawback that usually appears is the “intrusive” nature of certain changes [2], which means that it is difficult to accomplish seemingly “local” changes concerning certain entities without affecting others. The way one local change affects other parts of the ontology varies from the fact of blocking accesses to the ontology document that contains the object to be changed until the modification has been carried out, to changes concerning how one entity relates to others.

Cooperative systems are intrinsically dynamic. New tasks are continuously created and new members are incorporated into groups. Actors stop playing certain roles in order to start playing others just a click away. These facts are also defined in the cooperative system ontology. It is easy to see that some of the previously enumerated changes are more likely to occur than others, and some changes may be qualified as more specific for a particular cooperative system than other ones. The declaration of a new cooperative task such as “Mortgage Granting” is more general and therefore applicable in different systems than the declaration of the actor “Donald Johnson” as a new “Valuer”, which is only applicable to the “Valuation Office” he works for.

At the same time, technology enhancement enables people distributed geographically to cooperate in order to achieve common goals. Most current cooperation processes take place in the scope of Internet. Context-awareness about the changes in the shared environment is essential for successful cooperation [13]. Although the use and reuse of different sources in an ontology construction has been

widely advocated, it must be considered that an ontology which is too partitioned in multiple documents might also lead to maintenance troubles and might increase the latency of the system in retrieving documents. A compromise between efficiency, modularity and complexity has been looked for.

4.3 Behavior

An actor playing the role “Head of Risk” may play the role “Bank Manager” when he/she is absent. On the other hand, different actors (a notary, a valuer, a head of risk), working for different companies are engaged in the same group and the same cooperative task of granting a mortgage. Figure 5 shows one partial graphical representation of the mortgage system ontology written in OWL. Some of these relations have been highlighted using thick links. Behavioral aspects are represented within the same frame of the system ontology. Structural and behavioral aspects have been reflected by instantiating the ontology described in Figure 4.

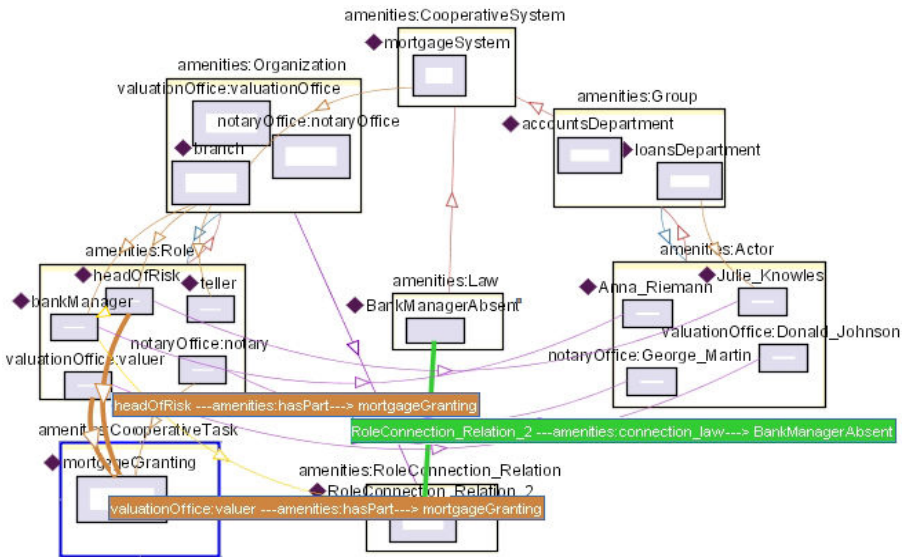


Fig. 5. Graphical representation of the branch cooperative system. Figure generated using the Jambalaya plug-in [17] for Protégé [20].

5 Conclusions

The starting point for any cooperation process between enterprises is a shared domain vocabulary. Conceptual frameworks provide analysts with the concepts, relations and the necessary terms to set the basis for a common understanding with stakeholders during the business process. They therefore constitute an appropriate starting point which is necessary so that domain ontologies may be formally defined. Domain ontologies are in turn the background for the definition of application ontologies. We

have followed this pattern to build ontologies that model workgroups of cooperative systems.

We have also proposed a three-tier design for ontology construction that preserves the distinction between domain and application ontologies and helps address the changes of the system ontology (which includes both the domain and the application ontology) in a less intrusive manner. While the first level corresponds to a domain ontology for cooperative systems, the second and third levels correspond to specific cooperative systems. The entities described in each one depend on their degree of abstraction. This representation scheme allows creating a generic, reusable model of enterprise business process. Furthermore, this proposal is rooted in organization and activities models and has the following characteristics:

- provides a shared terminology for the enterprises that each agent can jointly understand and use,
- defines the meaning of each term (group, law, role...) in a precise, “validatable” and unambiguous manner,
- implements the semantics that will enable to automatically deduce the answer to many questions about the organization and behavior of the enterprise and
- provides a graphical representation for depicting the enterprise business process.

Acknowledgements

This research is supported by R+D projects of the Spanish MCYT under project TIN2004-08000-C03-02.

References

1. Ambler, S. W., Nalbone, J., Vizdos, M.: Enterprise Unified Process: Extending the Rational Unified Process. Prentice Hall PTR, 2003.
2. Andrade, L.F., Fiadeiro, J. L.: Agility through Coordination. *Information Systems*, 27 (2002) 411-424
3. Andreasen, T., Nilsson, J.F.: Grammatical Specification of Domain Ontologies. *Data & Knowledge Engineering*, 48 (2004) 221-230
4. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: *The Description Logic Handbook*. Cambridge University Press, 2003
5. Chen, P., Akoka, J., Kangassalo, A., Thalheim, B.: *Conceptual Modeling: Current Issues and Future Directions*. LNCS 1565, 4-19, 1998.
6. Corcho, Ó., Fernández-López, M., Gómez-Pérez, A.: Methodologies, Tools and Languages for Building Ontologies: Where is their Meeting Point? *Data & Knowledge Engineering*, 46 (2003) 41-64
7. Evermann, J., Wand, Y.: Toward Formalizing Domain Modeling Semantics in Language Syntax. *IEEE Transactions on Software Engineering*, vol. 31, no. 1, January 2005
8. Fox, M.S., Barbeceanu, M., Gruninger, M.: An Organization Ontology for Enterprise Modelling: Preliminary Concepts for Linking Structure and Behaviour. *Wetice*, p. 71, 4th Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET-ICE'95), (1995)

9. Garrido, J.L., Gea, M., Rodríguez, M.L.: Requirements Engineering in Cooperative Systems. Requirements Engineering for Sociotechnical Systems. Chapter XIV, IDEA GROUP, Inc.USA (2005)
10. Greenberg, S.: Computer-Supported Cooperative Work and Groupware. Academic Press Ltd., London, UK (1991)
11. Gruninger, M., Atefi, K., and Fox, M.S.: Ontologies to Support Process Integration in Enterprise Engineering. Computational and Mathematical Organization Theory, Vol. 6, No. 4, pp. 381-394 (2000)
12. Guarino, N.: Formal ontology and information systems. In N. Guarino, Ed., Proceedings of FOIS '98, (Trento, Italy, June, 1998). IOS Press, Amsterdam, 1998, 3–15
13. Gutwin, C., Penner, R., Schneider, K.: Group Awareness in Distributed Software Development. Proc. ACM CSCW'04, 72-81 (2004)
14. Henderson, P., Crouch, S. and Walters, R. J.: Information Invasion in Enterprise Systems: Modelling, Simulating and Analysing System-level Information Propagation. In Proceedings of The 6th International Conference on Enterprise Information Systems (ICEIS 2004) 1, 473-481, Porto, Portugal. Seruca, I., Filipe, J., Hammoudi, S. and Cordeiro, J., Eds (2004)
15. Hurtado, M.V., Parets, J. Evolutionary Information and Decision Support Systems: An Integration Based on Ontologies. LNCS 2178, 146-159, 2001
16. Jaekel, F.W., Perry, N., Campos, C., Mertins, K., Chalmeta, R.: Interoperability Supported by Enterprise Modelling. LNCS 3762, 552 – 561, 2005
17. Jambalaya 2.2.0 build 15 2005/07/07 15:04. The Jambalaya Project. More info at <http://www.thechiselgroup.org/jambalaya>
18. Keijzer, A. de, Keulen, M. van: Information Integration - the process of integration, evolution and versioning, Technical report, December 2005, no. 05-58, Centre for Telematics and Information Technology (CTIT), ISSN 1381-3625
19. Kishore, R., Zhang, H. & Ramesh, R.: A Helix-Spindle Model for Ontological Engineering. Communications of ACM, February 2004/Vol. 47, No. 2, 69-75
20. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. LNCS 3298, 229 – 243, 2004
21. Noguera, M., González, M., Garrido, J.L., Hurtado, M.V., Rodríguez, M.L. System Modeling for Systematic Development of Groupware Applications. Proc. of the 5th International Workshop on System/Software Architectures, Las Vegas (USA), CSREA 2006
22. Olivé, A., Raventós, R.: Modeling Events as Entities in Object-Oriented Conceptual Modeling Languages. Data & Knowledge Engineering, xxx (2005) xxx. (Article in Press)
23. OWL Web Ontology Language Guide, Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, W3C Recommendation, 10 February 2004, <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>. Latest version available at <http://www.w3.org/TR/owl-guide/>
24. UML 2.0 Superstructure Specification (OMG), August 2003. Ptc/03-08-02, 455-510

MONET 2006 PC Co-chairs' Message

In recent years the research area of social mobile and networking technologies has made rapid progress, due to the increasing development of new mobile technologies and the widespread usage of the Internet as a new platform for social interactions.

Social applications of mobile and networking technologies serve groups of people in shared activities, in particular geographically dispersed groups who are collaborating on some task in a shared context. An important characteristic of these social applications is the continuous interaction between people and technology to achieve a common purpose. Again, social applications tend to be large-scale and complex, involving difficult social and policy issues such as those related to privacy and security access.

Mobile technologies are devoted to playing an important role in many areas of social activities, most likely in those areas where the right data at the right time have mission-critical importance. Mobile technologies play an essential role in personalizing working and interaction contexts, and supporting experimentation and innovation.

Social networking technologies join friends, family members, co-workers and other social communities together. These technologies are convergent, emerging from a variety of applications such as search engines and employee evaluation routines while running on equally diverse platforms from server clusters to wireless phone networks.

The third generation of social networking technologies has hit the Web. This network serves increasingly significant social functions. Networking technologies have to face emerging problems of robustness, such as vulnerabilities to reliability and performance due to malicious attack.

The first international workshop on MOBILE and NETWORKING TECHNOLOGIES for social applications (MONET 2006) was held on October 29, 2006 in Montpellier. MONET 2006 received 28 submissions, and from those the Program Committee selected the 13 papers in these proceedings. Each paper was reviewed by at least three referees and all evaluations were very convergent. We wish to thank, in particular, the Program Committee members who completed their reviews by the deadline, giving useful and detailed suggestions for improving papers.

The success of the MONET 2006 Workshop would not have been possible without the high level of organization of the OnTheMove (OTM) Federated Conferences.

August 2006

Fernando Ferri, National Research Council, Italy
Maurizio Rafanelli, National Research Council, Italy
Arianna D'Ulizia, National Research Council, Italy

Exploring Social Context with the Wireless Rope

Tom Nicolai¹, Eiko Yoneki², Nils Behrens¹, and Holger Kenn¹

¹ TZI Wearable Computing Lab, Universität Bremen, Germany
{nicolai, psi, kenn}@tzi.de

² University of Cambridge, UK
eiko.yoneki@cl.cam.ac.uk

Abstract. The Wireless Rope is a framework to study the notion of social context and the detection of social situations by Bluetooth proximity detection with consumer devices and its effects on group dynamics. Users can interact through a GUI with members of an existing group or form a new group. Connection information is collected by stationary tracking devices and a connection map of all participants can be obtained via the web. Besides interaction with familiar persons, the Wireless Rope also includes strange persons to provide a rich representation of the surrounding social situation. This paper seeks to substantiate the notion of *social context* by an exploratory analysis of interpersonal proximity data collected during a computer conference. Two feature functions are presented that indicate typical situations in this setting.

1 Introduction

As the field of wireless and locative technologies matures, a more enduring relationship between the physical and cultural elements and its digital topographies will become interesting topics to explore. Their interaction, influence, disruption, expansion and integration with the social and material practices of our public spaces will be getting more focus. Is public space a crowd of individuals? How can the crowd inspire the individual through collaboration, competition, confrontation? How change, effect, or experience could only be achieved by a mass movement, a cooperative crowd? How can we stage a series of new happenings? In [1], Huggle project takes an experiment of human mobility, where mobility gives rise to local connection opportunities when access infrastructure is not available. Our project Wireless Rope aims to take a further look from a social perspective.¹

Context awareness in general is recognized as an important factor for the success of ubiquitous computing applications and devices. The relevance of social context in particular was also noted, including the identities and roles of nearby persons (e.g. co-worker or manager) as well as the social situation [2].

¹ <http://wrp.auriga.wearlab.de>

Several works picked up the concept of sensing identities and used this information to annotate meeting recordings with a list of attendants [3] or to facilitate information exchange [4].

However, less is known about the recognition of the broader social situation on the basis of proximity data. This paper undertakes an initial exploration in the detection of such situations. The focus is on social contexts that do not presume knowledge about the identities and roles of individuals. For the approach presented here, it is not necessary to recognize the particular identities of individuals in the proximity. Instead it is interesting, e.g. if the person is with the others, or just passing them by, and if they are encountered regularly or not. This paper introduces two feature functions of proximity data to recognize several situations during a visit to a computer conference. Situations like arrival and departure, as well as coffee breaks and lunch are identifiable by this method.

With a robust classification of social contexts, an application would be able to detect meaningful episodes for a user while moving in different social circles and circumstances. Knowledge about these episodes could in turn be used to automatically adapt input and output modalities of a device (e.g. silent mode for mobile phones), to trigger actions (e.g. checking the bus schedule), or to guide the creation of an automatic diary according to episodes.

The paper is organized as follows: after a review of related work, the concept of proximity detection is elaborated. In section 4, the definition of the familiar stranger is given and its relevance to the classification of social situations is explained. The next section introduces the various components of the Wireless Rope system that was used to carry out the experiment described in section 6. The analysis of data and its discussion follows. The paper concludes with section 9.

2 Related Work

Social context has many different sides. At a very coarse level, it is related to the milieu a person lives in. Kurvinen and Oulasvirta examine the concept from a social science perspective [5]. They conclude, that the recognition of “turns” in activities gives valuable clues for an interpretation of social context. They also state that sensor data can only be interpreted for this purpose in the light of a well-defined domain.

Bluetooth proximity detection was already used in a number of other projects. Most notably, Eagle and Pentland used it to measure the social network of students and staff on a university campus in an extended experiment with one hundred students over the course of nine months [6]. Hui et al. carried out a similar study during a conference with the goal to identify prospects for ad-hoc networking scenarios [1]. Paulos and Goodman on the other hand use proximity detection to measure variables that might indicate the comfort in public urban places [12].

Proximity detection can also be realized by a number of other technologies. GPS can be used to capture the absolute position of two persons. A proximity

service with knowledge of both positions can then calculate the exact distance [7]. Infrared systems were already used in smart badges to detect people facing each other at conferences [8]. The Hummingbird system uses radio frequency to determine an approximate proximity in the range of 100m radius [9].

3 Proximity Detection with Bluetooth

The Wireless Rope uses Bluetooth for proximity detection. This technology is widely available and a lot of people carry a Bluetooth enabled mobile phone with them. Thus, it is possible to detect a certain amount of peoples' phones without handing a special device to each of them, which makes Bluetooth appealing for experiments involving a large quantity of persons.

The range of Bluetooth varies between 10m and 100m, depending on the device class. In mobile phones, the range is usually 10m. A part of the Bluetooth protocol stack is the *device inquiry*. It enables a device to discover other devices in the proximity—usually to establish a connection for data transfer. The discovery process requires active participation of the peer device. It may automatically answer an inquiry request or not, which can be configured by the user with the Bluetooth visibility option. If it answers, it discloses its device address and device class among others. The address uniquely identifies a Bluetooth device and can be used to recognize a formerly discovered device. The device class distinguishes mobile phones from computers and others and gives vague information about the further capabilities of a device.

The device inquiry does not give details about the distance to the device, except that it is in the communication range (i.e. 10m for most mobile phones). The measurement of the distance within the range is only possible indirectly by taking the bit error rate into account [10]. Unfortunately, additional software is necessary on the side of the discovered device, and a connection must be established prior to the measurement, which involves interaction by the user of the discovered device. Thus, the Wireless Rope uses the plain device inquiry mechanism to detect the proximity of other devices. It uses the device class to distinguish mobile phones from other devices to identify the proximity to other persons. The assumption here is, that the presence of a mobile phone indicates the presence of its owner. Mobile phones are very personal objects and are seldom left behind.

4 Familiar Strangers

To carry out a categorization of different social situations, some knowledge about the social structure of our modern lives is required. For the analysis presented here, the distinction between familiar and unfamiliar persons is important in particular.

Beyond this bipartite view, a third kind of social relationship emerged at the transition between familiar and strange persons with the urbanization of society: the familiar stranger. The sociologist Milgram did initial experiments regarding

this concept [11]. His definition of a familiar stranger is that it is person who is encountered repeatedly, but never interacted with. Typically, familiar strangers are encountered on the bus during ones daily way to work or while visiting the same recreational facilities. Paulos and Goodman presented a concept to recognize these persons with a device [12]. They state that such a device could be used to indicate the comfort a person feels in specific urban places.

Following this concept, we use a simple algorithm to distinguish strange persons from familiar strangers on the basis of proximity data. For the purpose of this paper, a familiar stranger must have been met more than five times. Different meetings are separated by periods of at least five minutes of absence. No further distinction between familiar strangers and familiar persons is considered here, although Eagle and Pentland remark that it could even be possible to identify friends on the basis of Bluetooth proximity data [6].

5 The Wireless Rope

To experiment with the notion of social context, we implemented a couple of components incorporating proximity detection. The Wireless Rope is a program for Java phones that collects information of surrounding devices using Bluetooth. It enables a group to actually feel the boundaries of the group. Like a real rope tying together mountaineers, the Wireless Rope gives the urban exploration group immediate feedback (tactile or audio) when a member gets lost or approaches. Thus everybody can fully engage in the interaction with the environment, and cognitive resources for keeping track of the group are freed.

Besides the direct interaction with familiar persons, the program also includes strangers and familiar strangers and recognizes them when they are met



Fig. 1. Sightings on phone display

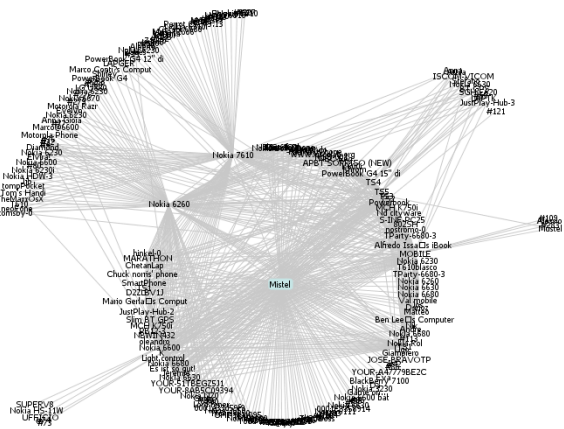


Fig. 2. Connection map on website

repeatedly. A glance at the program screen tells different parameters of the surrounding social situation: How many familiar and how many strange persons are in the proximity? How long have these persons been in proximity? Is there somebody with me for some time whom i have not noticed?

As an additional service, the collected information kept in all Wireless Rope programs may be gathered at a central server via special Track Stations. Users can look at the connection map created by gathered information from phones via the web (Fig. 2). The following subsections give details about the various components of the Wireless Rope.

5.1 Wireless Rope Program on Java Bluetooth Phones

The Wireless Rope program can be installed on mobile phones with Bluetooth that support the Java MIDP 2.0 and JSR-82 (Bluetooth) APIs. It performs periodic Bluetooth device inquiries to collect sightings of surrounding Bluetooth devices. Devices are classified into one of four categories and visualized as circles in different colors on the display:

Stranger (gray): All new sightings are classified as strangers.

Familiar Stranger (blue): Strangers which are sighted repeatedly by the proximity sensor are automatically advanced to the familiar stranger category.

Familiar (yellow): If the user recognizes a familiar person on the display, he can manually add him to the familiar category.

Contact (green): During an interaction with a person, both might agree to add themselves to their contacts (bidirectional link). Besides being notified of their proximity, contacts can use the Track Stations to exchange additional data.

While a device is in proximity the corresponding circle slowly moves from the top of the screen to the bottom. A time scale on the display lets the user interpret the positions of the circles. Proximity data are kept in the device until the information can be transmitted to a nearby Track Station.

5.2 Bluetooth Devices Without Wireless Rope

All Bluetooth devices that run in visible mode (respond to inquiries) are automatically included in the Wireless Rope and their sightings are collected. Users are notified of their existence and they are visualized on the display. The only difference is that these devices can not be added to the “Contact” category, because it involves a bidirectional agreement that is only possible with the Wireless Rope program.

5.3 Track Stations

Track Stations might be installed as additional infrastructure at highly frequented or otherwise meaningful locations, e.g. in conference rooms, train stations or bars. They consist of small Bluetooth enabled PCs in a box. The Track

Stations automatically record the passing-by of users by Bluetooth device inquiries and can transmit relevant digital tracks to contacts at a later time. They can notify trusted contacts of the last time they were seen by the station. By connecting these devices to the Internet, users can also check at which station a contact was seen the last time. By correlating the list of familiar strangers with the list of persons that often visit a station a user may see how much a place is “his kind of place.” Paulos and Goodman call this value “turf” [12]. Thus the Track Stations augment the reach of the Wireless Rope at important places. Periodically, these devices collect all log data from the mobile phones and aggregate them in a database for visualization and further analysis.

5.4 Reference Points

For roughly localizing the Wireless Rope users in space and to recognize a formerly visited place, reference points are used. Any stationary Bluetooth device can be used for this purpose. The Bluetooth device class is used to determine whether a device is stationary or not. The Bluetooth address then identifies a place.

5.5 Connection Map

The information collected by the Track Stations is visualized in realtime on a website. This connection map is anonymized for non-registered users. Registered users can explore their own neighbourhood including contacts, regularly met familiar strangers and randomly encountered strangers. The connection map is a tool for personal social network analysis, e.g. to identify common contacts and distinct cliques.

6 Experiment

The Wireless Rope was used to carry out an experiment to gather real-world proximity data for an exploratory analysis. The program was installed on a Nokia 6630 mobile phone to perform periodic Bluetooth device inquiries every 30 seconds.

The Ubicomp conference 2005 in Tokyo together with the workshop “Metapolis and Urban Life” was selected as a social event for the experiment for its varied program schedule, and because it was expected that a large proportion of the conference attendees had a detectable Bluetooth device with them. One of the attendants was carrying a prepared device during the entire time of the conference to collect the data. Additionally, he took photographs with the same device to document his activities. The program schedule of the conference provides detailed information about the planned timing of activities.

Since a significant amount of the encountered peoples’ phones was configured to answer these inquiries, it was possible to detect other phones and thus the related owners in a proximity of approximately ten meters. The data was recorded in the phone memory and later transferred to a computer for analysis.

The experiment ran over six days. On day one and two, the workshop took place. Part of the first day was an exploration of the city in the afternoon. Day three to five were spent on the main conference. The last day was spent with recreational activities in the city.

7 Data Analysis

The Wireless Rope provided the data used for the later analysis. Each device inquiry returned a set of unique device identifiers and additional information about the class of the devices. This data was recorded along with timestamps. The device class was used to filter out non-personal devices, like laptops and network equipment. In the next step, a set of quantitative features was extracted from the sets of device identifiers by a sliding time window of five minutes.

The features are chosen to be independent of the percentage of people that can be identified by the device inquiries. The proportion might change from situation to situation, with the particular mentalities of the people, cultural differences, and the general Bluetooth penetration in a country among others. Some groups of people are more extrovert than others and enable their Bluetooth visibility on purpose. Others are not aware about the consequences and might have it enabled randomly. Without independence from these factor, a comparison of data from different situations is difficult.

Let F_t be the set of all detected familiar persons in the time interval $[t, t + 1]$, and S_t the set of strangers respectively. For this experiment, only familiar and unfamiliar persons are distinguished. The familiar strangers are treated as being familiar.

The number of arriving familiar devices is $f_t^+ = |F_t| - |F_t \cap F_{t-1}|$ and $f_t^- = |F_{t-1}| - |F_t \cap F_{t-1}|$ is the number of leaving familiar devices. s_t^+ and s_t^- are defined correspondingly. The analyzed features indicate the dynamic in the group of familiars and strangers. They show how much an individual moves in accordance with the surrounding people:

1. $DynFam(t) = \frac{(f_t^+ + f_t^-) - ||F_t| - |F_{t-1}||}{|F_t|}$
2. $DynStra(t) = \frac{(s_t^+ + s_t^-) - ||S_t| - |S_{t-1}||}{|S_t|}$

8 Results and Discussion

The data set comprises 52411 Bluetooth sightings and 1661 meetings in total. Figure 3 and 4 show the histograms of individual Bluetooth sightings and derived meetings, respectively. There were approximately 650 registered conference visitors. 69 devices were classified as familiar and a total of 290 as strangers for the whole data set including conference and city encounters.

Figure 5 shows the features DynFam and DynStra for the six days of the experiment. The peaks indicate the different social activities the test subject was engaged in. The conference activity shows up clearly in the data. Arrival is

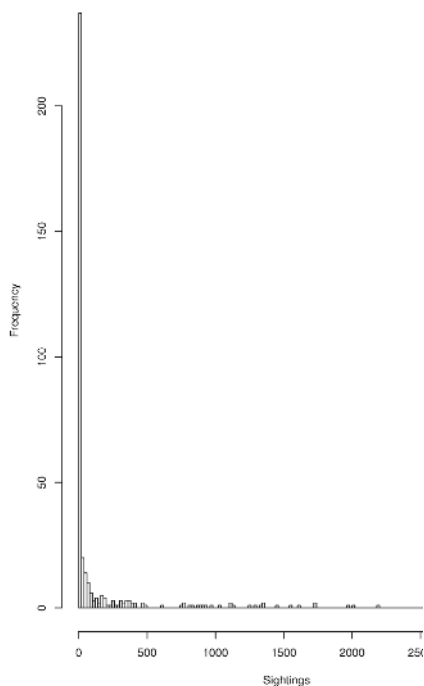


Fig. 3. Histogram of sightings

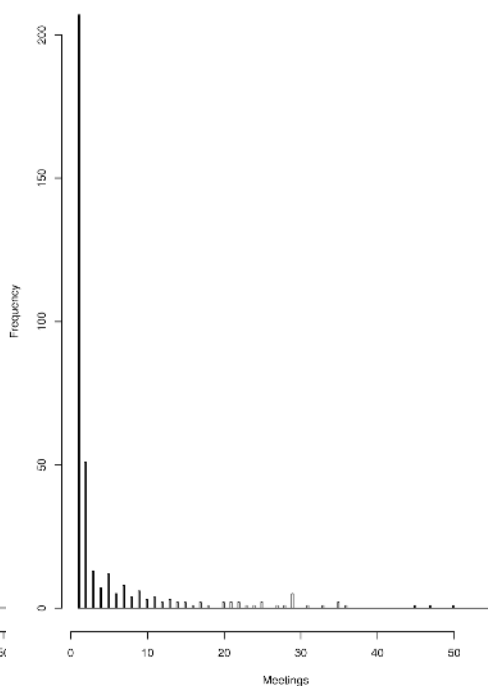


Fig. 4. Histogram of meetings

indicated by a peak in DynStra that is triggered during the movement through the crowded city. Coffee breaks, lunch and visits to the exhibition are indicated by peaks in DynFam. The workshop during day one and two is not detected, since the group behavior was rather homogeneous and did not exhibit the measured dynamic. The city exploration as part of the workshop on the other hand is clearly indicated. The arrival to the workshop did not require movement through crowds.

The peaks vary in width and height. The height relates to the frequency of the changing of people in the surrounding and the width to the duration of the changing. With the knowledge of the larger context—the conference visit in this case—it is possible to assign meanings to the individual peaks.

There were a couple of problems encountered with this experiment. First, Bluetooth is generally unpopular in Japan. Anyhow, most times there was enough reception in the city for this analysis. Only the movement in the night was not detected, although there were strangers on the streets. Inaccuracies in Bluetooth device inquiry were also discovered, but seem to have no significant negative effect (compare [6]). Moreover, the processing could not have been carried out like this during the measurement. The reason is, that the familiarity was calculated over the whole conference time before the features were calculated. Thus, effects of the process of getting familiar are not addressed here.

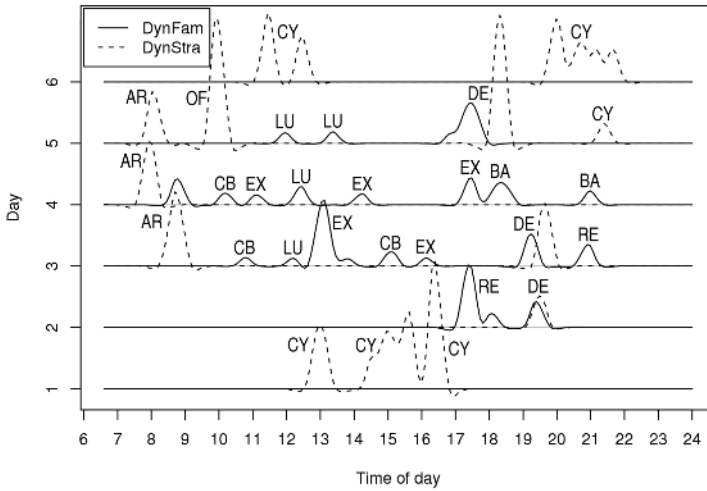


Fig. 5. Feature data of six days in Tokyo (smoothed by splines). Day 1 and 2: Workshop; day 3, 4, 5: Conference; day 6: day off. The peaks indicate social events or situations the test subject attended. CY: Moving through the city, RE: Conference reception, DE: Departure from conference, AR: Arrival at conference, CB: Coffee break, LU: Lunch, EX: Exhibition (posters and demos), BA: Banquet, OF: Off the conference.

9 Conclusion and Future Work

The Wireless Rope system was presented as a framework to experiment with proximity data in a variety of situations. It runs on modern mobile phones and collects proximity data by Bluetooth device inquiries. The analysis of data from a computer conference suggests, that the presented features are suited to indicate situations with a high dynamic in the movement of surrounding people on the basis of data collected by Bluetooth device inquiries. While movement in the city could also be detected by cheap location tracking technologies [13], the detection of movement within a building would require an expensive additional infrastructure. Even if other methods were in place, the classification of familiars and strangers in the proximity adds valuable information.

The conference was a well suited setting, since there was contact with a lot of different persons. Social relations are not very differentiated in this situation, since most persons are strangers at the beginning. The familiarity classifier indicates mainly, if someone is a regular conference attendee or not. In daily routine, a detailed discrimination of social roles, like family, friends and working colleagues would help to identify meaningful situations and episodes. As an alternative to the personal inquiry device, stationary devices could be used to measure the quality of a conference, e.g. to measure if sessions start on time, how popular individual sessions are, or how masses of people move through the conference space.

To further study this topic, it is necessary to determine the significance of these findings by comparing them to other persons, places, and scenarios. More features need to be developed and tested to account for other situations. Further, this method could be used in combination with other context sensors, like location. Correlation with a calendar could also yield interesting results. A learning algorithm could probably be used to determine the usual daily routine of a person and automatically detect meaningful deviations.

References

1. Hui, P., Chaintreau, A., Scott, J., Gass, R., Crowcroft, J., Diot, C.: Pocket switched networks and human mobility in conference environments. In: Proc. SIGCOMM 2005 Workshop on Delay Tolerant Networking, Philadelphia, USA, ACM Press (2005)
2. Schilit, B.N., Adams, N.I., Want, R.: Context-aware computing applications. In: Proc. Workshop on Mobile Computing Systems and Applications, Santa Cruz, USA, IEEE Computer Society (1994) 85–90
3. Kern, N., Schiele, B., Junker, H., Lukowicz, P., Tröster, G.: Wearable sensing to annotate meeting recordings. *Personal and Ubiquitous Computing* **7** (2003) 263–274
4. Kortuem, G., Segall, Z.: Wearable communities: Augmenting social networks with wearable computers. *IEEE Pervasive Computing* **2** (2003) 71–78
5. Kurvinen, E., Oulasvirta, A.: Towards socially aware pervasive computing: A turntaking approach. In: Proc. International Conference on Pervasive Computing and Communications (PerCom), Orlando, Florida, IEEE Computer Society (2004) 346–351
6. Eagle, N., Pentland, A.: Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing* **10** (2006) 255–268
7. Olofsson, S., Carlsson, V., Sjölander, J.: The friend locator: Supporting visitors at large-scale events. *Personal and Ubiquitous Computing* **10** (2006) 84–89
8. Gips, J., Pentland, A.: Mapping human networks. In: Proc. International Conference on Pervasive Computing and Communications (PerCom), Pisa, Italy, IEEE Computer Society (2006) 159–168
9. Holmquist, L.E., Falk, J., Wigström, J.: Supporting group collaboration with interpersonal awareness devices. *Personal Technologies* **3** (1999) 13–21
10. Madhavapeddy, A., Tse, A.: A study of bluetooth propagation using accurate indoor location mapping. In: Proc. Ubiquitous Computing (UbiComp), Tokyo, Japan, Springer Verlag (2005) 105–122
11. Milgram, S.: *The Individual in a Social World: Essays and Experiments*. Addison-Wesley (1977)
12. Paulos, E., Goodman, E.: The familiar stranger: Anxiety, comfort and play in public places. In: Proc. SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, ACM Press (2004) 223–230
13. Hightower, J., Consolvo, S., LaMarca, A., Smith, I., Hughes, J.: Learning and recognizing the places we go. In: Proc. Ubiquitous Computing (UbiComp), Tokyo, Japan, Springer Verlag (2005) 105–122

Extending Social Networks with Implicit Human-Human Interaction

Tim Clerckx, Geert Houben, Kris Luyten, and Karin Coninx

Hasselt University, Expertise Centre for Digital Media - IBBT,
and transnationale Universiteit Limburg
Wetenschapspark 2, BE-3590 Diepenbeek, Belgium
{tim.clerckx, geert.houben, kris.luyten,
karin.coninx}@uhasselt.be

Abstract. This paper describes a framework to enable implicit interaction between mobile users in order to establish and maintain social networks according to the preferences and needs of each individual. A user model is proposed which can be constructed by the user and appended with information regarding the user's privacy preferences. Design choices and tool support regarding the framework are discussed.

1 Introduction

The evolving world of Ubiquitous Computing [12] brings along quite a few new research areas such as mobile and context-aware computing, and new techniques to interact with the ubiquitous computing environment. Schmidt defined implicit interaction [11] as a new interaction model where interaction is based on information gathered by external sensory context information such as sound and location. The author describes examples where implicit interaction can be useful to interact with computers. However, this interaction model can be used similar to make humans communicate with each other through implicit interaction. We define this as *implicit human-human interaction (HHI)*.

Implicit HHI can be realized when each user carries a computing device with enabled wireless communication (cell phone, PDA, embedded device...) that contains a user profile covering information about the user and a statement of which information can be distributed to other users in the vicinity. That way, each user can point out which information may be provided to others. The other way around, the user can also describe what kind of information about other users can be useful in order to be notified when an *interesting* individual comes nearby.

Why can implicit HHI be useful? Increasing popularity of dating sites (50% increase in 2005 in the U.S.A. [2]) all over the internet reveals a market is available for the extension of social networks driven by profile matching. The next step would be to port this concept to the mobile device. Dodgeball [1], for instance, enables users to extend their social networks with people in their physical environment using a cell phone. Nonetheless other applications can be thought of. Consider the following scenario. A student is writing a paper about his project but he is having some difficulties on finding

related work. This is why he indicates on his PDA he wishes to make contact to a person who is a field expert on the subject. When the student encounters such a person, the expert will receive a message. Later that day, the expert decides to contact the student in order to help him with his problem. In this way, the student got in touch with a tutor without explicitly contacting the individual. This scenario will be used in the examples further down this paper.

Making implicit HHI possible requires some choices to be made regarding technical issues:

- some kind(s) of **wireless communication** has to be used;
- an appropriate **user model** to represent the user has to be constructed;
- the user should be able to easily edit his/her preferences regarding the user profile and the security settings about these preferences (**privacy**);
- an **agent** has to represent the user to decide which incoming information is useful to the user and which local information may be released to other agents;

The remainder of this paper will discuss how we built a framework where these issues are tackled in order to enable implicit HHI. First we will discuss some work related to the goals we wish to reach (section 2). Afterwards we will define the user model we have used to describe the user (section 3.1), the privacy related to the user's information (section 3.2) and the link between these two models (section 3.3). Section 4 discusses an overview of the framework; and tools we have implemented to enable implicit HHI are described. Finally future work will be discussed and conclusions will be drawn.

2 Related Work

As discussed in the previous section, implicit HHI can be realized when data about a user is emitted on one side and interpreted and matched to user data on another side. Ferscha et al. [5] recognized this kind of bidirectional communication between entities (humans or objects equipped with a computing device) and defined the interaction model called *Digital Aura* (because of the metaphorical resemblance with the energy field surrounding a person). They based their model on the *Focus and Nimbus* model [10]: nimbus describes the kind of information about the entity that is observable, while focus describes the kind of information that is observed by the entity. As a result the intersection of focus and nimbus equals the interaction.

Cavalluzzi et al. [4] constructed a multi-agent system called D-Me (Digital-Me). The name reveals a personal agent is a representative for the user in taking decisions in active environments. The agent interacts with other agents in the user's direct vicinity in order to decide which tasks are currently available to be performed in the vicinity according to the user's goals. An application is discussed in [3] where users interact with a digital travelling guide. An agent deduces interesting information about the environment on the basis of a user profile and decides which information will be delivered to the user and how it will be presented. The user can also provide feedback which the agent will use in order to properly adjust the user's goals.

3 User Model and Privacy

Systems that respond to the user's goals and preferences require notion about the user. For a system to make use of its knowledge about the user, this information has to be structured and stored in a *user model*.

A user model suitable for implicit HHI should not only contain the user's profile and preferences but also information about the privacy sensitiveness related to this data. This is why the user model (\mathcal{M}) is divided in these two parts that also make up the structure of this section together with the link between these two parts: the user profile (\mathcal{U}) and the privacy profile (\mathcal{P}):

$$\mathcal{M} = \mathcal{U} \cup \mathcal{P}$$

3.1 A User Profile for Implicit Human-Human Interaction

The use of user models and profiles in interactive systems is not a new concept. Kobsa [8] discusses historical and recent developments in the field of user models. In order to construct a suitable user profile for implicit HHI, we looked at several existing techniques for modelling the user. GUMO (General User Model Ontology [6]), for instance, is an ontology which can be used to model concepts about the real world, such as users, to be used in ubiquitous computing applications. To represent the user in the implicit HHI framework, we use some concepts of GUMO. To introduce the privacy we need to enable implicit HHI, we will extend this with the privacy profile (section 3.2).

The user profile, as in GUMO, consists of a finite set of subjects consisting of situational statements. The user can choose to divide the profile into distinct subjects regarding the user, for instance: identity, work, love and free time.

$$\mathcal{U} = \{U_{\text{identity}}, U_{\text{work}}, U_{\text{love}}, U_{\text{freetime}}\}$$

We divide the user profile into distinct social categories at this level because they comprise a first categorization required for the privacy (as we will discuss later on in this work). This is an advantage because when a user chooses subjects like in the example, these are already related to social interaction. Identity contains information such as the user's name, address, date of birth, etc. The three other subjects are related to the origin of interaction with other people: information about the user's employment, the user's marital status, the way the user spends his/her free time. . . Each subject U in \mathcal{U} is a collection of situational statements containing information about the subject:

- information regarding the statement:
 - the *auxiliary* (a): the relation between the predicate and the subject;
 - the *predicate* (p): the constituent of the statement;
 - the *range* of values (r) possible for the situational statement;
- information regarding the statement's value:
 - the actual *value* (v);
 - the *start* (s) time of the current value;
 - the *duration* (d) describing how long the current value is valid;
 - the level of *trust* ($t \in \{\text{None}, \text{Low}, \text{Medium}, \text{High}\}$) describing the confidentiality of the statement's information.

In short:

$$\forall U \in \mathcal{U} : U = \bigcup_i \{(a_i, p_i, r_i) \longrightarrow (v_i, s_i, d_i, t_i)\}$$

For instance consider the following valid situational statement related to the scenario described in the introduction. Here a field expert reveals to anyone ($t_i = \text{None}$) he has an excellent knowledge about UbiComp and his level of excellence will last for at least a year.

$$U_{\text{work}} \supseteq \{(\text{hasKnowledge}, \text{UbiComp}, \text{TerribleMediocreGoodExcellent}) \longrightarrow (\text{Excellent}, 4/05/2006\ 19 : 04 : 29, \text{year}, \text{None})\}$$

3.2 Representing Privacy in the User Model

The user profile described in the previous section has to be enriched with information to ensure the user’s privacy (with respect to other *human* users). Because privacy depends on several factors like the user’s liking, situation, or location; the user has to be given the opportunity to indicate which data about him/herself is visible to which people. This is why we introduce the *privacy profile* to describe the user’s preferences about the privacy of his/her personal information.

The privacy profile is divided into privacy categories the same way as the user profile. For example when we take the same categories as in the previous section:

$$\mathcal{P} = \{P_{\text{identity}}, P_{\text{work}}, P_{\text{love}}, P_{\text{freetime}}\}$$

Each category in the privacy profile consists of two sets: a set of groups (G) that collects contacts of the user and a set of context statements (C). Each group of contacts (g) describes which humans (h) belong to the group and the level of trust ($t \in \{\text{None}, \text{Low}, \text{Medium}, \text{High}\}$) assigned to this group for the current privacy category describing how trustworthy the group of humans is. The set of context statements is a collection of contextual situations (c) describing location and/or temporal information where the trust level ($t \in \{\text{Never}, \text{None}, \text{Low}, \text{Medium}, \text{High}\}$) will overrule the confidentiality level described in the user profile. This is summarized in the following expression:

$$\forall P \in \mathcal{P} : P = \{G, C\} = \left\{ \bigcup_i \{(g_i, t_{g_i})\}, \bigcup_j \{(c_j, t_{c_j})\} \right\} : g_i = \{h_1, \dots, h_n\}$$

Consider the following example. A user deems colleagues as highly trustworthy considering work-related information and friends as less trustworthy. Furthermore when the user is on holiday, he does not wish to share work-related information to any group. This can be represented by the following statements:

$$P_{\text{work}} = \{G_{\text{work}}, C_{\text{work}}\}, G_{\text{work}} \supseteq \{(\text{colleagues}, \text{High}), (\text{friends}, \text{Low})\}, \\ C_{\text{work}} \supseteq \{(\text{onHoliday}, \text{Never})\}$$

Dividing the user's contacts into distinct groups has proven to be a good technique in information disclosure to other people (e.g. the Precision Dial interaction framework of Lederer [9]). In our approach, security levels are not restricted to the assignment of security levels to contact groups but they are the result of several dependencies: categorization of content in the user profile (identity, work...), categorization of contacts (friends, colleagues...), and external context information (contextual statements) as we will discuss in the next section.

3.3 Linking the User and Privacy Profile

The user profile \mathcal{U} and privacy profile \mathcal{P} now have to be combined in order to decide which information can be distributed among which users. In order to decide this, a link between the profiles is necessary. This link was implicitly established in the categorization of the situational statements in \mathcal{U} and the use of the same categorization in \mathcal{P} . Let us clarify this by means of an example. Consider the example of the situational statement in section 3.1. The user expresses his knowledge about UbiComp and declares this information is not confidential at all. Because this statement belongs to the *Work* category in \mathcal{U} , the privacy statements of the *Work* category in \mathcal{P} will reflect on this piece of user data. As a result anyone can access the information because the user has indicated the user data to be of the lowest confidentiality level. If the user suddenly decides to change the confidentiality level of a statement s in the user profile \mathcal{U} to $t_s = \text{Medium}$, only contacts belonging to the group of *colleagues* will be able to access the user data according to the example statements in section 3.2 ($t_s = \text{Medium} \leq \text{High} = t_{\text{colleagues}}$ and $t_s = \text{Medium} > \text{Low} = t_{\text{friends}}$).

This can be generalized in the following expression describing which statements in the user profile \mathcal{U} should be released to a human h in the user's vicinity:

$$\begin{aligned} \forall U_i \in \mathcal{U}, \forall s = (a_s, p_s, r_s) \longrightarrow (v_s, s_s, d_s, t_s) \in U_i : \text{release}(s, h) \Leftrightarrow \\ \exists P_j \in \mathcal{P} : i = j, \exists G_j \in P_j, \exists (g_k, t_{g_k}) \in G_j : h \in g_k \wedge t_s \leq t_{g_k} \end{aligned}$$

In short this means the confidentiality level of the user data (t_s) has to be less or equal than the trustworthy level of the group the receiving human belongs to (t_{g_k}).

In section 3.2 we discussed the possibility to involve external context information in the privacy profile. It is possible to specify a context statement that describes a situation where the regular confidentiality levels specified in the user profile are overruled. In the example the user described to overrule the work-related privacy settings when the user is on holiday. This implies that the confidentiality level of all work-related user data is increased to the maximal level. As a result, when a situation in conformance with a contextual statement c_j with confidentiality level t_{c_j} occurs ($(c_j, t_{c_j}) \in G_j \in P_j \in \mathcal{P}$), the previous expression to release information to a human h should be replaced by:

$$\forall U_i \in \mathcal{U} : i = j, \forall s = (a_s, p_s, r_s) \longrightarrow (v_s, s_s, d_s, t_s) \in U_i : \text{release}(s, h) \Leftrightarrow t_s \leq t_{c_j}$$

This means the confidentiality level of the user data (t_s) should be less or equal than the confidentiality level imposed by the context situation (t_{c_j}). Because the confidentiality level of a context situation can also be *Never*, it is possible to completely overrule every privacy level, and release no information about the user profile for a privacy category $P_j \in \mathcal{P}$.

4 Architecture

In this section we elaborate on the architecture of the framework to support implicit HHI. When two people carrying a mobile device are nearby each other, spontaneous interaction takes place. These local interactions are limited to the physical proximity of people because of the restrictions on the range of today’s wireless communication nodes. The amount of information exchanged between two entities relies on mutual similarity of their profiles.

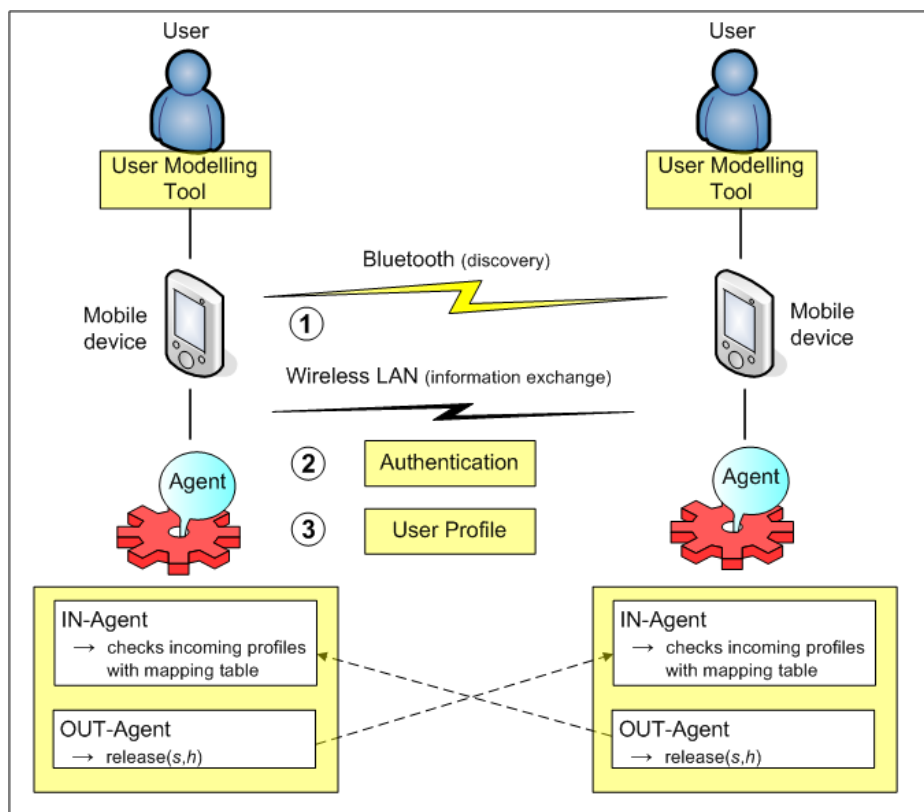


Fig. 1. Architectural overview of the implicit human-human interaction framework

Fig. 1 shows an overview of the architecture and the different modules of the framework. The proposed framework is modular and easily extensible; the different parts in the system can easily be modified or replaced to improve, update or adapt the application (e.g. to implement a different agent to interpret incoming user profiles). The framework consists of three core parts: a *communication* module, the *user modelling tool* and the autonomous *personal agent*. These modules will be described separately in the following paragraphs. The introduction (section 1) includes a concrete scenario to

point out a real life situation in which the adoption of the interaction model is demonstrated. The screenshots used in this section are based on the described scenario. A view on the main screen of the application is shown in Fig. 3(b). Centralized in the user interface is the message window, on which incoming messages are visualized. On top and at the bottom of the screen the user can adjust application settings.

4.1 Communication

Before two users can exchange personal information implicitly, discovering and authentication of the other party has to be accomplished. Peer-to-peer short range communication capabilities are needed to discover (step 1 in Fig. 1) other devices and to exchange information (step 2 and 3 in Fig. 1). At a first stage authentication information is exchanged in order to identify the other user. Bluetooth is used for discovery because the limitations of this technology regarding the short range ensure the detected users are in the user's direct proximity. Then the user's personal agent authenticates other users and decides which parts of the user profile can be shared with each particular user as described in section 3.3. The authentication and exchange stages are using WiFi in order to cope with users that get out of reach after detection.

4.2 User Modelling Tool

The personal user-related content within the framework and basic component of the interaction model is the user model, which covers the user profile and the privacy profile. In the growing digital world most users are sceptical and reserved regarding personal information sharing, especially in the implicit HHI context. To trust an information management system, users must have full control of the data and the privacy related to the data. In the first place, an orderly structured and easy to use user interface for editing a user model is required [7]. Especially when we are considering a user model that is going to be shared with other users.

Fig. 2(a) shows a screenshot of the user profile overview. The situational statements for each category are listed with the corresponding predicate, auxiliary, value and the level of trust (indicated by the color gradation). By clicking on an entry, the statement editor is shown and allows the user to change the different fields (Fig. 2(b)).

Not only controlling shared content is important, also the opportunity to adjust the privacy settings regarding this information is crucial. The privacy profile and the link with the user profile are defined in section 3.2 and 3.3. Although the personal agent model takes care of the in and outgoing messages, it is the user who is responsible for entering the correct settings regarding the release of data according to the appropriate context. Fig. 3(a) gives a concrete example of a security profile overview. The different entries are visualizations of the defined formalizations in section 3.

4.3 Personal Agent

The personal agent is an important part of the framework because it is responsible for spreading and examining the outgoing (*OUT-agent*) and incoming (*IN-agent*) user profiles, as shown in Fig. 1. These agents act autonomously and collaboratively to relieve the user from manually distributing parts of his/her user profile and from examining

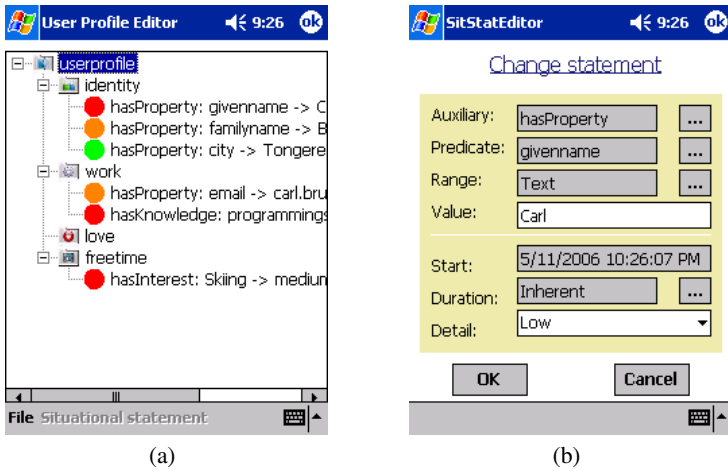


Fig. 2. (a) Overview interface of a user profile. (b) Situational statements editor.

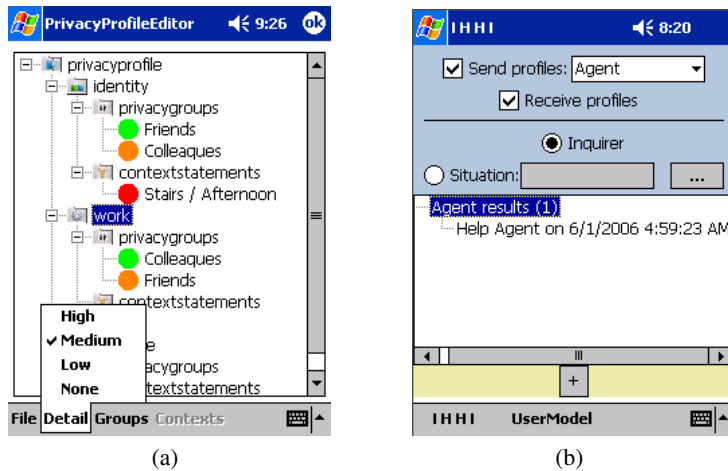


Fig. 3. (a) Overview interface of a security profile. (b) Main screen of the application.

other user profiles. There has to be a well defined and linear mapping between the content of the user and security information and the actions of the agent. The user has to be aware of its behavior to assure the right level of confidence when using the application.

After connecting with an entity in the vicinity, the OUT-agent creates a subset of the user profile, based on the identity of the other person and the privacy settings related to this person (the default trust level for an unknown person is None). To produce the subset of the user profile, the technique described in section 3.3 is used.

On the other hand the IN-agent checks the incoming user profile information from other users in the vicinity on similarities with the profile of his user to decide if there are common interests. To filter information relevant for the user, the incoming

situational statement is compared to the situational statement in the user's own profile using a mapping table which describes relevant matches. For instance, the statement (hasKnowledge, UbiComp, TerribleMediocreGoodExcellent) \rightarrow (Excellent, 4/05/2006 19 : 04 : 29, year, None) will match to (needsHelp, UbiComp, YesNo) \rightarrow (Yes, 4/05/2006 20 : 16 : 33, day, None). When a relevant match is found, the user is notified with a short message on his device (shown in Fig. 3(b)).

5 Discussion and Future Work

In this paper we discussed a framework to support implicit human-human interaction (HHI). The user can define a user profile and adjust privacy settings according his/her requirements. Due to the separation of the user profile and the privacy settings and because they are loosely coupled, it is possible to change the privacy settings instantly. Furthermore it is possible to automatically change the privacy settings depending on external context information. More information about the framework (such as movies and XML-serialization of the models) can be found on the website¹. In future work we plan field tests to explore how users react to implicit HHI when they are using applications built on our framework. Furthermore network security has to be implemented to ensure the concealment of the peer-to-peer communication.

Acknowledgments

The authors would like to thank Carl Bruninx for his contributions to this work. Carl has contributed significantly in the design and development of the work presented in this paper. Part of the research at EDM is funded by EFRO (European Fund for Regional Development), the Flemish Government and the Flemish Interdisciplinary institute for Broadband Technology (IBBT). The CoDAMoS (Context-Driven Adaptation of Mobile Services) project IWT 030320 is directly funded by the IWT (Flemish subsidy organization). The scenarios kept in mind for the research in this paper have considered several projects, such as the FP6 IP MYCAREVENT (IST nr. 004402). The MYCAREVENT project is an Integrated Project sponsored by the European Commission in support of the Strategic Objective "Information Society Technologies (IST)" in the Sixth Framework Program.

References

1. *dodgeball.com*. Google Inc., <http://www.dodgeball.com/>, 2006.
2. Rhys Blakely. Valentine hackers target lovelorn surfers. Times Online, February 08, 2006, <http://business.timesonline.co.uk/article/0,,9075-2030844,00.html>, 2006.
3. Addolorata Cavalluzzi, Berardina De Carolis, Sebastiano Pizzutilo, and Giovanni Cozzolongo. Interacting with embodied agents in public environments. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 240–243, New York, NY, USA, 2004. ACM Press.

¹ <http://research.edm.uhasselt.be/~tclerckx/ihhi/>

4. Giovanni Cozzolongo, Berardina De Carolis, and Sebastiano Pizzutilo. A Personal Agent Supporting Ubiquitous Interaction. In Matteo Baldoni, Flavio De Paoli, Alberto Martelli, and Andrea Omicini, editors, *WOA 2004: Dagli Oggetti agli Agenti. 5th AI*IA/TABOO Joint Workshop "From Objects to Agents": Complex Systems and Rational Agents, 30 November - 1 December 2004, Torino, Italy*, pages 55–61. Pitagora Editrice Bologna, 2004.
5. A. Ferscha, M. Hechinger, R. Mayrhofer, M. dos Santos Rocha, M. Franz, and R. Oberhauser. Digital Aura. In A. Ferscha, H. Hörtner, and G. Kotsis, editors, *Advances in Pervasive Computing*, volume 176, pages 405–410. Austrian Computer Society (OCG), April 2004. part of the Second International Conference on Pervasive Computing (Pervasive 2004).
6. Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. Gumo - The General User Model Ontology. In Liliana Ardissono, Paul Brna, and Antonija Mitrovic, editors, *User Modeling*, volume 3538 of *Lecture Notes in Computer Science*, pages 428–432. Springer, 2005.
7. Judy Kay. The UM Toolkit for Cooperative User Modeling. *User Modeling and User-Adapted Interaction*, 4(3):149–196, 1995.
8. Alfred Kobsa. Generic user modeling systems. *User Modeling and User-Adapted Interaction*, 11(1-2):49–63, 2001.
9. Scott Lederer. Designing Disclosure: Interactive Personal Privacy at the Dawn of Ubiquitous Computing. University of California at Berkeley, M.S. Report, Berkeley, CA, USA, 2003.
10. Tom Rodden. Populating the application: a model of awareness for cooperative applications. In *CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pages 87–96, New York, NY, USA, 1996. ACM Press.
11. Albrecht Schmidt. Implicit Human Computer Interaction Through Context. *Personal and Ubiquitous Computing*, 4(2/3), 2000.
12. Mark Weiser. The computer for the 21st century. *Scientific American*, 265(3):66–75, January 1991.

A Classification of Trust Systems

Sebastian Ries*, Jussi Kangasharju, and Max Mühlhäuser

Department of Computer Science
Darmstadt University of Technology
Hochschulstrasse 10
64289 Darmstadt, Germany

{ries, jussi, max}@tk.informatik.tu-darmstadt.de

Abstract. Trust is a promising research topic for social networks, since it is a basic component of our real-world social life. Yet, the transfer of the multi-facetted concept of trust to virtual social networks is an open challenge. In this paper we provide a survey and classification of established and upcoming trust systems, focusing on trust models. We introduce a set of criteria as basis of our analysis and show strengths and short-comings of the different approaches.

1 Introduction

Trust is a well-known concept in everyday life, which simplifies many complex processes. Some processes are just enabled by trust, since they would not be operable otherwise. On the one hand, trust in our social environment allows us to delegate tasks and decisions to an appropriate person. On the other hand, trust facilitates efficient rating of information presented by a trusted party. Computer scientists from many areas, e.g., security, ubiquitous computing, semantic web, and electronic commerce, are still working on the transfer of this concept, to their domain. In Sect. 2 we will introduce, the main properties of social trust, in Sect. 3 we provide our own set of criteria and the analysis of a selected set of trust systems from different areas, and in Sect. 4 we give a short summary and derive ideas for our future work.

2 Properties of Trust

There is much work on trust by sociologists, social psychologists, economists, and since a few years also by computer scientists. In general trust can be said to be based on personal experience with the interaction partner in the context of concern, on his reputation, or on recommendations. Furthermore, trust is connected to the presence of a notion of uncertainty, and trust depends on the expected risk associated with an interaction. [1, 2, 3, 4, 5, 16]

* The author's work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the PhD program "Enabling Technologies for Electronic Commerce" at Darmstadt University of Technology.

The following properties are regularly assigned to trust, and are relevant when transferring the concept to computer sciences. Trust is subjective and therefore asymmetric. It is context dependent, and it is dynamic, meaning it can increase with positive experience and decrease with negative experience or over time without any experience. This makes also clear that trust is non-monotonic and that there are several levels of trust including distrust. A sensitive aspect is the transitivity of trust. Assuming Alice trusts Bob and Bob trusts Charlie, what can be said about Alice trust in Charlie? In [2], Marsh points out that trust is not transitive. At least it is not transitive over arbitrary long chains, since this will end in conflicts regarding distrust. Yet recommendation and reputation are important factors for trust establishment.

McKnight and Chervany state in [1] that there are three principle categories of trust: personal / interpersonal trust, impersonal / structural trust, and dispositional trust. Interpersonal trust describes trust between people or groups. It is closely related to the experiences, which people had with each other. Structural trust is not bound to a person but raises from social or organizational situation. Dispositional trust can be explained as a person's general attitude towards the world. As shown in [6] much work is done on transferring interpersonal trust to computer sciences, whereas there is little work supporting the other categories.

Although, trust is a well-known concept and despite there is a set of properties on which most researchers agree, it is hard to define trust. A couple of definitions are provided from several scientific areas with different focuses and goals (cf. [2, 6]). A definition which is shared or at least adopted by some researchers [3, 7, 8, 9], is the definition provided by the sociologist Diego Gambetta:

"... trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent will perform a particular action, both before [we] can monitor such action (or independently of his capacity of ever to be able to monitor it) and in a context in which it affects [our] own action." [16]

3 Classification Criteria and Analysis

Having introduced the general aspects of trust, we will now give a survey how the concept of trust is realized in different areas of computer science. We derive our coarse-grained classification from the work provided in [2, 4, 5, 10, 11]. As main categories we see *trust modeling*, *trust management* and *decision making* [12]. In this classification, trust modeling deals with the representational and computational aspects of trust values. Trust management focuses on the collection of evidence and risk evaluation. Although decision making is actually a part of trust management, we treat it separately, since it is such an important aspect.

Due to the limitations of this paper and our own research interests we focus for a more fine-grained classification only on trust modeling, especially on the aspects of *domain*, *dimension*, and *semantics* of trust values.

Trust values are usually expressed as numbers or labels, thus their domain can be binary, discrete, or continuous. A binary representation of trust allows

only to express the two states of "trusted" and "untrusted". This is actually near to certificate- or credential-based access control approaches, where access is granted, if and only if the user presents the necessary credentials. But since most researchers agree that trust has several levels, binary models are considered as not sufficient. Trust can also be represented using more than two discrete values. This can be done either by using labels or by using a set of natural numbers. The advantage of this approach is, that trust values can be easily assigned and understood by human users [3, 13]. Continuous trust values are supported by well-known mathematical theories depending on the semantics of the trust values.

The dimension of trust values can be either one- or multi-dimensional. In one-dimensional approaches this value usually describes the degree of trust an agent assigns to another one, possibly bound to a specific context. Multi-dimensional approaches allow to introduce a notion of uncertainty of the trust value.

The semantics of trust values can be in the following set: rating, ranking, probability, belief, fuzzy value. As rating we interpret values which are directly linked with a trust related semantics, e.g., on a scale of natural numbers in the interval [1, 4], 1 can be linked to "very untrusted",..., and 4 to "very trusted". Whereas, the trust values which are computed in ranking based models, e.g., [14], are not directly associated with a meaningful semantics, but only in a relative way, i.e. a higher value means higher trustworthiness. Therefore, it is only possible to assign an absolute meaning to a value, if this value can be compared to large enough set of trust values of other users. Furthermore, trust can be modeled as probability. In this case, the trust value expresses the probability that an agent will behave expected. The details of belief and fuzzy semantics are explained together with 'Subjective Logic' and ReGreT (see below). A summary of our classification is presented in Table 1.

Table 1. Classification of trust models

	Domain	Dimension	Sem.	Trust management	Decision making
Marsh	cont. in [-1,1)	1 (situational trust)	rating	- (but risk evaluation)	threshold-based
TidalTrust	disc. in [1, 10]	1 (rating)	rating	global policy (no risk evaluation)	-
Abdul-Rahman & Hailles	disc. labels	1 (trust value)	rating	-	-
SECURE Project (exemplary)	disc. in [0, ∞]	2 (evid.-based)	prob.	local policies (incl. risk evaluation)	threshold-based
	cont. in [0, 1]	3 (bel., disbel., uncert.)	belief		
Subjective Logic	disc. in [0, ∞]	2 (evid.-based)	prob.	not directly part of SL	not directly part of SL
	cont. in [0, 1]	3 (b, d, u)	belief		
ReGreT	disc. fuzzy values	2 (trust, confidence)	fuzzy values	local policies (fuzzy rules)	-

3.1 Model Proposed by Marsh

The work of Marsh [2] is said to be the seminal work on trust in computer science. Marsh concentrates on modeling trust between only two agents. He introduces knowledge, utility, importance, risk, and perceived competence as important aspects related to trust. The trust model should be able to answer the questions: With whom should an agent cooperate, when, and to which extend? The trust model uses real numbers in $[-1; 1)$ as trust values. He defined three types of trust for his model. Dispositional trust \mathcal{I}_x is trust of an agent x independent from the possible cooperation partner and the situation. The general trust $\mathcal{T}_x(y)$ describes the trust of x in y , but is not situation specific. At last, there is the situational trust $\mathcal{T}_x(y, a)$, which describes the trust of agent x in agent y in situation a . The situational trust is computed by the following linear equation:

$$\mathcal{T}_x(y, a) = \mathcal{U}_x(a) \times \mathcal{I}_x(a) \times \widehat{\mathcal{T}_x(y)} , \quad (1)$$

where $\mathcal{U}_x(a)$ represents the utility and $\mathcal{I}_x(a)$ the importance, which x assigns to the trust decision in situation a . Furthermore, $\widehat{\mathcal{T}_x(y)}$ represents the estimated general trust of x in y .

The trust management provided by Marsh does not treat the collection of recommendations provided by other agents, he only models direct trust between two agents. The aspect of risk is dealt with explicitly based on costs and benefits of the considered engagement.

The decision making is threshold based. Among other parameters the cooperation threshold depends on the perceived risk and competence of the possible interaction partner. If the situational trust is above the value calculated for the cooperation threshold, cooperation will take place otherwise not. Furthermore, the decision making can be extended by the concept of "reciprocity", i.e. if one does another one a favor, it is expected to compensate at some time.

3.2 TidalTrust

In [13] Golbeck provides a trust model which is based on 10 discrete trust values in the interval $[1, 10]$. Golbeck claims that humans are better in rating on a discrete scale than on a continuous one, e.g., in the real numbers of $[0, 1]$. The 10 discrete trust values should be enough to approximate continuous trust values. The trust model is evaluated in a social network called FilmTrust [15] with about 400 users. In this network the users have to rate movies. Furthermore, one can rate friends in the sense of "[...] if the person were to have rented a movie to watch, how likely it is that you would want to see that film" [13].

Recursive trust or rating propagation allows to infer the rating of movies by the ratings provided by friends. For a source s in a set of nodes S the rating r_{sm} inferred by s for the movie m is defined as

$$r_{sm} = \frac{\sum_{i \in S} t_{si} \cdot r_{im}}{\sum_{i \in S} t_{si}} , \quad (2)$$

where intermediate nodes are described by i , t_{si} describes the trust of s in i , and r_{im} is the rating of movie m assigned by i . To prevent arbitrary long recommendation chains, the maximal chain length or recursion depth can be limited. Based on the assumption that the opinion of the most trusted friends are the most similar to opinion of the source, it is also possible to restrict the set of considered ratings, to those provided by the most trusted friends.

Although the recommendation propagation is simple, the evaluation in [13] shows that it produces a relatively high accuracy, i.e. the ratings based on recommendation are close to the real ratings of the user. Since this approach does not deal with uncertainty, the calculated trust values can not benefit in case that there are multiple paths with the similar ratings. The trust value is calculated as a weighted sum. For the same reason, the path length does not influence the trust value. The values for trust in other agents on the path are used for multiplication and division in each step. Since each node aggregates its collected ratings and passes only a single value to its ancestor in the recursion, the source cannot evaluate which nodes provided their rating. The approach does not deal with any form of risk or decision making.

3.3 Model Proposed by Abdul-Rahman and Hailes

The trust model presented by Abdul-Rahman and Hailes [7] is developed for use in virtual communities with respect to electronic commerce and artificial autonomous agents. It deals with a human notion of trust as it is common in real world societies. The formal definition of trust is based on Gambetta [16].

The model deals with direct trust and recommender trust. Direct trust is the trust of an agent in another one based on direct experience, whereas recommender trust is the trust of an agent in the ability of another agent to provide good recommendations. The representation of the trust values is done by discrete labeled trust levels, namely "Very Trustworthy", "Trustworthy", "Untrustworthy" and, "Very Untrustworthy" for direct trust, and "Very good", "good", "bad" and, "very bad" for recommender trust.

A main aspect of this trust model is to overcome the problem that different agents may use the same label with a different subjective semantics. For example, if agent a labels an agent c to be "Trustworthy" based on personal experience, and a knows that agent b labels the same agent c to be "Very Trustworthy". The difference between these two labels can be computed as "semantic distance". This "semantic distance" can be used to adjust further recommendations of b .

Furthermore, the model deals with uncertainty. Uncertainty is introduced if an agent is not able to determine the direct trust in an agent uniquely, i.e. if an agent has e.g., as much "good" as "very good" experiences with another agent. But it seems unclear how to take benefit from this introduction of uncertainty in the further trust computation process. The combination of recommendations is done as weighted summation. The weights depend on the recommender trust and are assigned in an ad-hoc manner.

Although the model drops recommendations of unknown agents for the calculation of the recommended trust value, those agents get known by providing

recommendations, and their future recommendations will be used as part of the calculation.

It is important to mention that the direct trust values are only used to calculate the semantic distance to other agents, but are not used as evidence which could be combined with the recommendations.

Trust management aspects are not considered. The collection of evidence is only stated for recommendations of agents which have direct experience with the target agent. It is not explicitly described how to introduce recommendations of recommendations. Furthermore, the system does not deal with risk. Decision making seems to be threshold based, but is not explicitly treated.

3.4 SECURE Project Trust Model

The trust model and trust management in the SECURE project [5, 17] aims to transfer a human notion of trust to ubiquitous computing.

A main aspect of the trust model is to distinguish between situations in which a principal b is "unknown" to a principal a , and situations in which a principal b is "untrusted" or "distrusted". The principal b is unknown to a , if a cannot collect any information about b . Whereas b is "untrusted" if a has information, based on direct interaction or recommendations, stating that b is an "untrustworthy" principal.

This leads to define two orderings on a set of trust values \mathcal{T} denoted as \preceq and \sqsubseteq . The first ordering (\mathcal{T}, \preceq) is a complete lattice. For $X, Y \in \mathcal{T}$ the relation $X \preceq Y$ can be interpreted as Y is more trustworthy than X . The second ordering $(\mathcal{T}, \sqsubseteq)$ is a complete partial order with a bottom element. The relation $X \sqsubseteq Y$ can be interpreted as the trust value Y is based on more information than X .

The set of trust values can be chosen from different domains as long as the orderings have the properties described above. It is possible to use intervals over the real numbers in $[0, 1]$ [17]. This allows for an interval $[d_0, d_1]$ to introduce the semantics of belief theory by defining d_0 as belief and $1 - d_1$ as disbelief. Uncertainty can be defined as $d_1 - d_0$. Another possibility would be to define the trust values as pair of non-negative integers (m, n) . In this case m represents the number of non-negative outcomes of an interaction and n the number of negative ones. These approaches seem to be similar to the trust model provided by Jøsang, but they do not provide a mapping between these two representations. It is also possible to define other trust values e.g., discrete labels.

The trust propagation is based on policies. This allows users to explicitly express whose recommendations are considered in a trust decision. Let \mathcal{P} be the set of principals, the policy of a principal $a \in \mathcal{P}$ is π_a . The local policy allows to assign trust values to other agents directly, to delegate the assignment to another agent, or a combination of both. Since it is possible to delegate the calculation of trust values, the policies can be mutually recursive. The collection of all local policies π can be seen as global trust function m . This function m can be calculated as the least fixpoint of Π , where Π is $\Pi : \lambda p : \mathcal{P}. \pi_p$.

The trust management also deals with the evaluation of risk. Risk is modeled based on general cost probability density functions, which can be parameterized

by the estimated trustworthiness of the possible interaction partner. The evaluation of risk can be based on different risk policies, which e.g., describe if the risk is independent from the costs associated to an interaction or if it increases with increasing costs.

The decision making is threshold based. For the application in an electronic purse [5] two thresholds are defined by the parameters x, y ($x \leq y$). If the situation specific risk value (parameterized by the trust value corresponding to the interaction partner) is below x , the interaction will be performed (money will be payed), if it is above y the interaction will be declined. In case the risk value is between x and y the decision will be passed to the user.

3.5 Subjective Logic

The trust model presented by Jøsang [10], named "subjective logic", combines elements of Bayesian probability theory with belief theory. The Bayesian approach is based on beta probability density function (pdf), which allows to calculate posteriori probability estimates of binary events based on a priori collected evidence. For simplification we do not explain the concept of atomicity, which is introduced by Jøsang to use his model also for non-binary events.

The beta probability density function f of a probability variable p can be described using the two parameters α, β as:

$$f(p \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1 - p)^{\beta-1}, \tag{3}$$

where $0 \leq p \leq 1, \alpha > 0, \beta > 0$.

By defining $\alpha = r + 1$ and $\beta = s + 1$, it is possible to relate the pdf directly to the priori collected evidence, where r and s represent the number of positive and negative evidence, respectively. In this model trust is represented by opinions which can be used to express the subjective probability that an agent will behave as expected in the next encounter. It is possible to express opinions about other agents and about the truth of arbitrary propositions. The advantage of this model is that opinions can be easily be derived from the collected evidence.

An approach to deal with uncertainty is called belief theory, which tempts to model a human notion of belief. In belief theory as introduced in [10] an opinion can be expressed as a triple (b, d, u) , where b represents the belief, d the disbelief, and u the uncertainty about a certain statement. The three parameters are interrelated by the equation $b + d + u = 1$. Jøsang provides a mapping between the Bayesian approach and the belief approach by defining the following equations:

$$b = \frac{r}{r + s + 2}, \quad d = \frac{s}{r + s + 2}, \quad u = \frac{2}{r + s + 2} \quad \text{where } u \neq 0 \text{ .} \tag{4}$$

Furthermore, he defines operators for combining (consensus) and recommending (discounting) opinions. In contrast to the belief model presented in [18] the consensus operator is not based on Dempster's rule. Moreover, the model supports also operators for propositional conjunction, disjunction and negation.

In [19] it is shown how "subjective logic" can be used to model trust in the binding between keys and their owners in public key infrastructures. Other papers introduce how to use "subjective logic" for trust-based decision making in electronic commerce [20] and how the approach can be integrated in policy based trust management [21].

Another approach modeling trust based on Bayesian probability theory is presented by Mui et al. in [8], an approach based on belief theory is presented by Yu and Singh in [18].

3.6 ReGreT

ReGreT tries to model trust for small and mid-size environments in electronic commerce [22]. The system is described in detail in [23, 24]. A main aspect of ReGreT is to include information which is available from social relations between the interacting parties and their environments. In the considered environment the relation between agents can be described as competitive (*comp*), cooperative (*coop*), or trading (*trd*).

The model deals with three dimensions of trust or reputation. The individual dimension is based on self-made experiences of an agent. The trust values are called direct trust or outcome reputation. The social dimension is based on third party information (witness reputation), the social relationships between agents (neighborhood reputation), and the social role of the agents (system reputation). The ontological dimension helps to transfer trust information between related contexts. For all trust values a measurement of reliability is introduced, which depends on the number of past experience and expected experience (*intimate* level of interaction), and the variability of the ratings.

The trust model uses trust or reputation values in the range of real numbers in $[-1; 1]$. Overlapping subintervals are mapped by membership functions to fuzzy set values, like "very good", which implicitly introduce semantics to the trust values. In contrast to the probabilistic models and belief models, trust is formally not treated as subjective probability that an agent will behave as expected in the next encounter, but the interpretation of a fuzzy value like "very good" is up to the user or agent.

Since the fuzzy values are allowed to overlap, this introduces also a notion of uncertainty, because an agent can be e.g., "good" and "very good" at the same time to a certain degree.

The inference of trustworthiness is based on intuitively interpretable fuzzy rules. The trustworthiness assigned by agent a to agent c with respect to providing information about agent b , e.g., can depend on the relation between the agents b and c , as shown in the following example. In the example the social trust of a in information of b about c is "very bad" if the cooperation between b and c is high.

IF $coop(b; c)$ is *high*
THEN $socialTrust(a; b; c)$ is *very bad*.

Further information concerning risk evaluation and decision making is not given.

4 Conclusion

In this paper we have provided a short survey of trust systems based on different approaches. Furthermore, we provided a set of criteria to analyze systems dealing with trust, on a top level by distinguishing between trust model, trust management and decision making, and for the main aspects of trust modeling in detail. As we can see from our survey, it is possible to reason about trust models without especially addressing aspects of trust management, and the other way around. The comparison of trust models is yet difficult, since they are often developed for different purposes and use different semantics for modeling trust. Furthermore, most authors define their own way of trust management to evaluate their trust models. The trust propagation chosen by Golbeck seems to be a simple and yet an accurate way to evaluate recommendations in social networks.

By analyzing the trust models, we came to the conclusion that the models need to be able to represent a notion of uncertainty or confidence, since it is a main aspect of trust. The approach taken in ReGreT allows to define a subjective component for confidence, but the approach seems to be done in an ad hoc manner. The approach taken by belief models binds uncertainty to belief and disbelief. In conjunction with the Bayesian approach uncertainty depends directly of the number of collected evidence, but it is not related to a subjective and context-dependent measurement. For our future work we favor the Bayesian approach, since it allows to easily integrate the collected evidence. We will try to find a new way to derive uncertainty from the relation between the amount of collected evidence and an amount of expected evidence based on this approach. By giving the user the opportunity to define an expected amount of evidence, uncertainty gets a subjective and most notably a context-dependent notion.

References

1. McKnight, D.H., Chervany, N.L.: The meanings of trust. Technical report, Management Information Systems Research Center, University of Minnesota (1996)
2. Marsh, S.: Formalising Trust as a Computational Concept. PhD thesis, University of Stirling (1994)
3. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. In: Decision Support Systems. (2005)
4. Grandison, T., Sloman, M.: A survey of trust in internet applications. IEEE Communications Surveys and Tutorials **3**(4) (2000)
5. Cahill, V., et al.: Using trust for secure collaboration in uncertain environments. IEEE Pervasive Computing **2/3** (2003) 52–61
6. Abdul-Rahman, A.: A Framework for Decentralised Trust Reasoning. PhD thesis, University College London (2004)
7. Abdul-Rahman, A., Hailes, S.: Supporting trust in virtual communities. In: Proc. of Hawaii International Conference on System Sciences. (2000)
8. Mui, L., Mohtashemi, M., Halberstadt, A.: A computational model of trust and reputation for e-businesses. In: Proc. of the 35th Annual HICSS - Volume 7, Washington, DC, USA, IEEE Computer Society (2002)

9. Teacy, W.T., et al.: Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems* **12**(2) (2006)
10. Jøsang, A.: A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **9**(3) (2001) 279–212
11. Grandison, T., Sloman, M.: Specifying and analysing trust for internet applications. In: *I3E '02: Proc. of the IFIP Conference on Towards The Knowledge Society*, Deventer, The Netherlands, Kluwer, B.V. (2002) 145–157
12. Ries, S.: Engineering Trust in Ubiquitous Computing. In: *Proc. of Workshop on Software Engineering Challenges for Ubiquitous Computing*, Lancaster, UK (2006)
13. Golbeck, J.: Computing and Applying Trust in Web-Based Social Networks. PhD thesis, University of Maryland, College Park (2005)
14. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in p2p networks. In: *Proc. of the 12th international conference on World Wide Web*, New York, USA, ACM Press (2003) 640–651
15. Golbeck, J., Hendler, J.: Filmtrust: Movie recommendations using trust in web-based social networks. In: *Proc. of the Consumer Communications and Networking Conference*. (2006)
16. Gambetta, D.: Can we trust trust? In Gambetta, D., ed.: *Trust: Making and Breaking Cooperative Relations*. Basil Blackwell, New York (1990) 213–237
17. Carbone, M., Nielsen, M., Sassone, V.: A formal model for trust in dynamic networks. In: *Proc. of IEEE International Conference on Software Engineering and Formal Methods*, Brisbane, Australia, IEEE Computer Society (2003)
18. Yu, B., Singh, M.P.: An evidential model of distributed reputation management. In: *Proc. of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY, USA, ACM Press (2002) 294–301
19. Jøsang, A.: An algebra for assessing trust in certification chains. In: *Proc. of the Network and Distributed System Security Symposium*, San Diego, USA, (1999)
20. Jøsang, A.: Trust-based decision making for electronic transactions. In: *Proc. of the 4th Nordic Workshop on Secure IT Systems*, Stockholm, Sweden (1999)
21. Jøsang, A., Gollmann, D., Au, R.: A method for access authorisation through delegation networks. In: *4th Australasian Information Security Workshop (Network Security) (AISW 2006)*. Volume 54 of CRPIT., Hobart, Australia, ACS (2006)
22. Sabater, J., Sierra, C.: Review on computational trust and reputation models. *Artificial Intelligence Review* **24**(1) (2005) 33–60
23. Sabater, J., Sierra, C.: Reputation and social network analysis in multi-agent systems. In: *Proc. of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems*, New York, NY, USA, ACM Press (2002) 475–482
24. Sabater, J.: Trust and reputation for agent societies. PhD thesis, Institut d'Investigacion en Intelligencia Artificial, Spain (2003)

Solving Ambiguities for Sketch-Based Interaction in Mobile Environments

Danilo Avola, Maria Chiara Caschera, and Patrizia Grifoni

Istituto di Ricerche sulla Popolazione e Politiche Sociali,
CNR, Via Nizza 128, 00198 Rome, Italy

{danilo.avola, mc.caschera, patrizia.grifoni}@irpps.cnr.it

Abstract. The diffusion of mobile devices and the development of their services and applications are connected with the possibility to communicate anytime and anywhere according to a natural approach, which combines different modalities (speech, sketch, etc.). A natural communication approach, such as sketch-based interaction, frequently produces ambiguities. Ambiguities can arise in sketch recognition process by the gap between the user's intention and the system interpretation.

This paper presents a classification of meaningful ambiguities in sketch-based interaction and discusses methods to solve them taking into account of the spatial and temporal information that characterise the drawing process. The proposed solution methods use both sketch-based approaches and/or integrated approaches with other modalities. They are classified in: prevention, a-posteriori and approximation methods.

Keywords: Sketch based interaction, ambiguity resolution, mobile environments.

1 Introduction

Mobile devices are evolved from simple tools of communication to multifunctional one, able to support each human and social activity. For these reasons the use of mobile devices with different modalities such as speech, gesture, sketch, and so on [1],[2],[3] can be particularly relevant. Using mobile devices can make communication intuitive and spontaneous. However the naturalness can produce ambiguities. The sketch activity supports a natural communication approach and it allows to the user to convey easily simple or complex input/output on mobile devices, however it is intrinsically ambiguous [4], [5].

The purpose of this paper is to analyse the sketch-based interaction and its ambiguities. An ambiguity produces a semantic gap between communicative user's intention and its interpretation. The interpretation of the user's sketch involves the understanding of the informative content of a generic performed sketch. The informative content is expressed by spatial and temporal information of the sketch activity. Some interesting discussions about ambiguities in sketch based interaction are presented in [6], [7], [8] and [9].

This paper provides a classification of ambiguities for sketch-based interaction and a classification of their solutions methods. Ambiguities are classified as: i) ambiguities due to crosses in a stroke, ii) ambiguities due to the over-tracing of different

strokes, iii) ambiguities due to the intersection of two polygons, iv) ambiguities due to the inaccuracy of the user's tracing, and v) ambiguities due to the deleting and re-tracing actions. Solution methods are grouped in: i) prevention methods of ambiguities, ii) a-posteriori resolution methods, iii) approximation resolution methods.

The paper is organized as follows: section 2 proposes the classification of the ambiguities, Section 3 discusses the methods to solve these ambiguities and, finally, Section 4 concludes.

2 Classification of Ambiguities

Many kinds of ambiguities arise during the sketch's interpretation. This section introduces a classification that concerns with ambiguities due to; i) crosses in a stroke, ii) over-tracing of different strokes, iii) intersection of two polygons, iv) inaccuracy of the user's tracing and v) deleting and re-tracing actions (stroke is a drawing action).

Before dealing with ambiguities it is necessary to introduce some concepts. During the sketch process two elementary actions are considered: drawing a stroke and deleting an area. A drawing action is spatially characterized by the trajectory starting when the pen-tip begins to touch the tablet/paper and ending at the time it leaves the tablet/paper. A deleting action is spatially characterized by an area. These actions are temporally characterized by the temporal interval in which they are performed and by the velocity of the gesture for each point of the trajectory. Because of a user could introduce discontinuities in correspondence to the angles during the drawing activities of polygons or polylines, it is important to consider any spatial and/or temporal discontinuity that appears during the drawing actions. According to the spatial and temporal discontinuity the generalised stroke concept is introduced. It considers strokes modified by deleting actions and any spatial and/or temporal discontinuity during the drawing.

Below ambiguities are organized in different classes underlying their main features.

2.1 Ambiguities Due to Crosses in a Stroke

When the user draws one or more than one symbol by one stroke only (pen down, pen movement, pen up sequence) ambiguities can arise. They can be due to crosses in the stroke. In fact, some configurations can have more than one interpretation. For example, let us consider the sketch of Fig. 1a. It is composed by one stroke σ_a only and the user has not performed deleting actions. Considering spatial and temporal discontinuity, three generalized strokes σ'_1 , σ'_2 and σ'_3 are produced by the cross point (Fig. 1b). Three different interpretations are possible for Fig. 1a: the first one consists of one polygon and two polylines (Fig. 1c), the second consists of one polygon and one polyline (Fig. 1d) and, finally, the third is given by one polyline only (Fig. 1e).

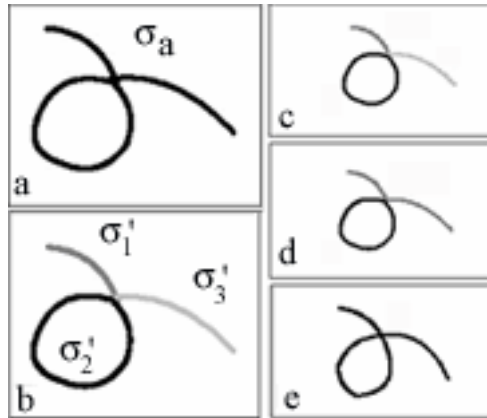


Fig. 1. Ambiguity due to crosses in a stroke

2.2 Ambiguities Due to the Over-Tracing of Different Strokes

Unlike the previous case, which involves one stroke only, this section introduces a further class of ambiguities that appear when the user over-traces two or more than two different strokes. When the user traces a sequence of pixels over a different one, spatially contiguous pixels can present a temporal discontinuity. Fig. 2 shows two over-traced strokes in order to explain this kind of ambiguities. The user has not performed deleting actions.

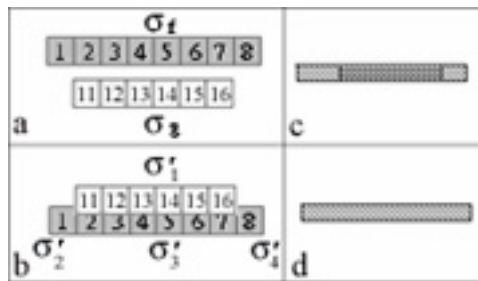


Fig. 2. Ambiguity due to Over-tracing of sequences of pixels belonging to two different strokes

In particular Fig. 2a shows the two strokes according to a temporal point of view. The temporal range of values 1-8 characterizes the stroke σ_f . The second stroke σ_g (Fig. 2a) over-traces the stroke σ_f and the temporal range of values 11-16 characterizes it. For simplicity sake spatial overlapping is not represented in Fig. 2a. The same pixels are related to the temporal range of values 2 - 7 and to the temporal range of values 11 - 16. Pixels with temporal values equal to 1 and 11 are spatially contiguous but they present a temporal discontinuity. A similar situation can be observed for pixels with temporal values equal to 16 and 8. According to spatial and temporal discontinuity four generalized strokes σ'_1 , σ'_2 , σ'_3 and σ'_4 are considered in

Fig. 2b. Several interpretations of the four generalized strokes can be given such as, for example, two different over-traced strokes (Fig. 2c) or one stroke only (Fig. 2d).

2.3 Ambiguities Due to the Intersection of Two Polygons

Another class of ambiguities is produced by the intersection of two strokes. Let us draw two strokes σ_b and σ_c in sequence according with Fig. 3a and, let us suppose the user has not performed deleting actions. According to the spatial and temporal discontinuity four generalized strokes σ'_1 , σ'_2 , σ'_3 and σ'_4 can be considered (Fig. 3b). The sketch of Fig. 3a could have a set of different interpretations. For brevity sake only four of them are showed in (Fig. 3c). They identify: i) three different polygons A, B, and C ii) two overlapped polygons (Fig. 3d) iii) two polygons and one polyline (Fig. 3e) and iv) two polygons and one polyline (Fig. 3f).

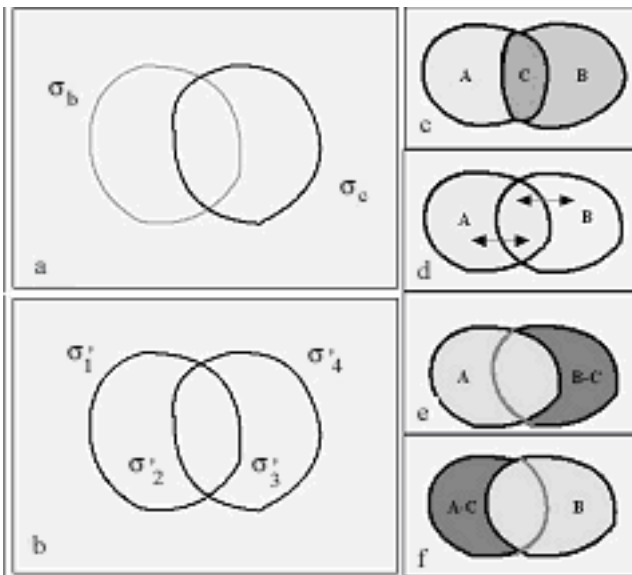


Fig. 3. Ambiguity due to the A and B polygons intersection

2.4 Ambiguities Due to the Inaccuracy of the User's Tracing

A further class of ambiguities can be given by inaccuracy of the user's tracing. This inaccuracy can produce more than one interpretation due to the gap between user's intention and how he/she is able to convey it.

The following sections show two different kinds of ambiguities due to inaccuracy: i) ambiguities due to the missing closure of a polygon, ii) ambiguities due to the generation of undesired polygons and polylines.

2.4.1 Ambiguities Due to the Missing Closure of a Polygon

A first class of ambiguities is observed when inaccuracy of the free-hand tracing can be interpreted either as a polyline or a non-closed polygon. Let us draw a stroke σ_d

according to Fig. 4a and, let us suppose the user has not performed deleting actions. Therefore, according to spatial and temporal discontinuity one generalized stroke σ_1 can be considered and it is the same of the σ_d . The sketch of Fig. 4a could have two different interpretations: i) a polygon (Fig. 4b), and ii) a polyline (Fig. 4c).

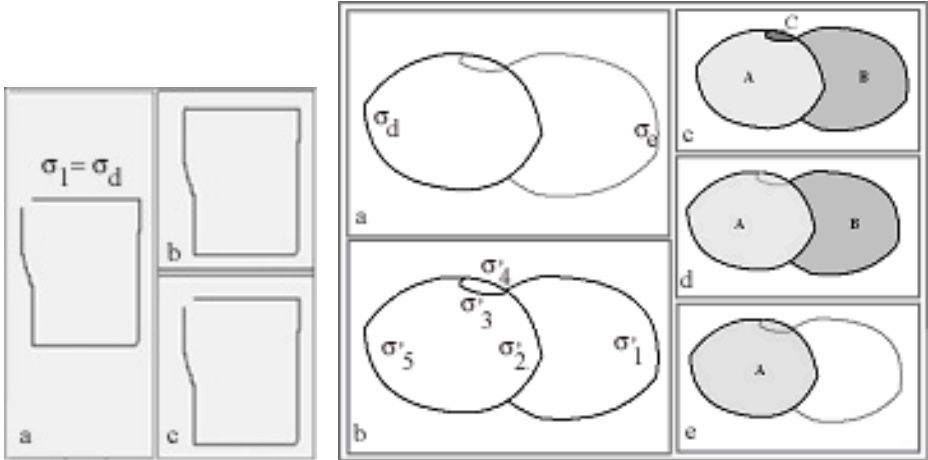


Fig. 4. Ambiguity due to the missing closure of a polygon **Fig. 5.** Ambiguity due to the generation of an undesired polygon

2.4.2 Ambiguities Due to the Generation of Undesired Polygons and Polylines

A different class of ambiguities is observed when inaccuracy produces undesired polygons. Let us consider two strokes σ_d and σ_e that represent the sketch. Let us suppose the second stroke produces a small third polygon due to inaccuracy in the user drawing and also in this case the user has not performed deleting actions (Fig. 5a). Five generalized strokes $\sigma'_1, \sigma'_2, \sigma'_3, \sigma'_4$ and σ'_5 are considered (Fig. 5b) according to spatial and temporal discontinuity. The sketch of Fig. 5a could have three different interpretations: i) three different polygons A, B, and C (Fig. 5c) ii) two polygons (Fig. 5d) iii) one polygon and one polyline (Fig. 5e).

In this case the most probably correct interpretation is shown in Fig. 5d, which considers two different polygons (A and B). However, the users could really desire to draw a small polygon in correspondence to the boundary of A and B. Similarly the inaccuracy can produce undesired small polylines that are not distinguishable by really desired small polygons.

2.5 Ambiguities Due to the Deleting and Re-tracing Actions

A new class of ambiguities is produced by deleting and re-drawing parts of a sketch. For example, given two different polygons, that are obtained by drawing two different strokes σ_h and σ_k , let us suppose the common boundary between them has been deleted and successively one or more than one stroke is drawn as a new boundary between the two polygons (Fig. 6a).

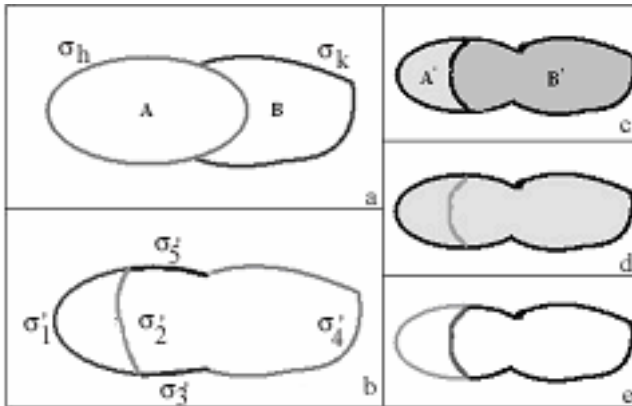


Fig. 6. Ambiguity produced by the deleting and re-tracing actions

Five generalized strokes σ'_1 , σ'_2 , σ'_3 , σ'_4 and σ'_5 are produced according to spatial and temporal discontinuity (Fig. 6b) consequently to the deleting action. Fig. 6a shows two polygons that have common boundary pixels between them. The common boundary can be deleted and new pixels are redrawn (Fig. 6b). The boundary of A' is intercepted by σ'_1 and σ'_2 . The boundary of B' is composed by σ'_2 , σ'_3 , σ'_4 and σ'_5 . The most used interpretation considers the two different polygons A' and B' (Fig. 6c). Other possible interpretations consider respectively a polygon and a polyline (Fig. 6d), or three polylines (Fig. 6e). Previously cited ambiguities can be generalized considering one or more than one polylines intercepting a polygon. They have at least one or more intersection points and one set of points that are neither internal points nor boundary points of the polygon.

3 Solution of Ambiguity

In this section three classes of methods for solving ambiguities are proposed. They are: i) prevention methods of ambiguities, ii) a-posteriori resolution methods, iii) approximation resolution methods. Below these methods and their main features are described focusing on their possibility of solving the previous classes of ambiguities. Table 1 provides a synthetic view of the different classes of ambiguities and their resolution methods, detailed in the following sections.

3.1 Prevention Methods of Ambiguities

The main method to prevent ambiguities is the procedural one. This method imposes to the user to respect a pre-defined interaction behaviour. This method is usually adopted in the Command User Interfaces, but it can be used in the Sketch-based Interfaces too.

In this case it can prevent several classes of ambiguities such as: ambiguities due to crosses in a stroke, ambiguities due to the intersection of two polygons, ambiguities due to the over-tracing of different strokes. Let us suppose, for example, that system

Table 1. Ambiguities and their solution methods

Methods		Ambiguities						
		Crosses in a stroke	Over-tracing of different strokes	Inaccuracy		Intersection of two polygons	Deleting and retracing actions	
				Missing closure of a polygon	Generation of undesired polygons and polylines			
Prevention methods of ambiguities	Procedural method	x	x			x		
A-posteriori resolution methods	Repetition	Modality	x	x	x	x	x	x
		Granularity of Repair			x	x		
	Choice	x	x	x	x	x	x	
	Choice/Beautification			x	x			
Approximation resolution methods	Thresholding			x	x			
	Historical Statistics		x	x	x			
	Rules	x	x	x	x	x	x	

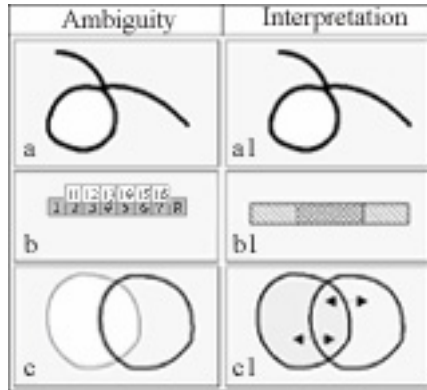


Fig. 7. Ambiguity solved by prevention method

imposes to the user the constraint to associate one and only one object to each stroke, where an object is a polyline or a polygon. According to this constraint the previous ambiguities can be solved in the following way: i) the sketch's ambiguity due to crosses in a stroke of Fig. 7a can be interpreted as the polyline shown in Fig. 7a₁, ii) the ambiguity due to the over-tracing of different strokes of Fig. 7b can be interpreted as the Fig. 7b₁ that represents two over-traced strokes, iii) the ambiguity of the Fig. 7c due to the intersection of two polygons can be interpreted as Fig. 7c₁ that represents two overlapped polygons.

3.2 A-Posteriori Resolution Methods of Ambiguities

The main method for the a-posteriori resolution of ambiguities is the mediation [9]. It consists of two sub-methods: repetition and choice. In repetition the user repeats an action until the system is able to interpret it correctly. In the choice method the system returns to the user the candidate set of interpretations and the user can select the more correct one.

Repetition can be divided in two sub-methods [9]: i) Modality method, ii) Granularity of Repair method. The first one focuses on the modality used in the repetition. It can involve one or more modalities. Involving more than one modality to solve ambiguities is more effective than the repetition by the same modality, because the user frequently replies the same errors and ambiguities when he/she uses the same modality. The need to solve ambiguities can imply the user has to add useful information for the disambiguation process. This information can be complementary, redundant or concurrent and it can be provided using different modalities. These different modalities can interact according to six basic types of cooperation [10] (Table 2).

Table 2. Types of cooperation between modalities

Types of cooperation between modalities	Description
Complementarity	different chunks of information composing the same command are transmitted over more than one mode
Equivalence	a chunk of information may be transmitted using more than one mode
Redundancy	the same chunk of information is transmitted using more than one mode
Transfer	a chunk of information produced by one mode is analysed by another mode
Concurrency	independent chunks of information are transmitted using different modalities and overlap in time
Specialization	a specific chunk of information is always transmitted using the same mode

When repetition combines different modalities redundancy is mainly used to solve ambiguities. When modalities are redundant, the system integrates the same chunk of information that is transmitted using more than one mode. This information can be jointly used to solve all the ambiguities introduced in the second section and an example of redundancy is provided in the following. Let us consider ambiguities due to the missing closure of a polygon (see Fig. 8a). In order to solve this kind of ambiguities, system could allow user to repeat the same input using a different modality, for example voice. Once user ends his/her drawing, he/she can say the word “polygon”. This information allows to the system to interpret the sketch as a polygon (Fig. 8a₁). Similarly to the previous case, ambiguities due to crosses in a stroke, ambiguities due to intersection of two polygons, ambiguities due to the generation of undesired polygons and polylines and, finally, ambiguities due to the over-tracing of different strokes can be solved using the repetition by another modality. Combining sketch based and speech-based interaction can be particularly useful on Mobile devices according to the interaction needs arising in the different contexts.

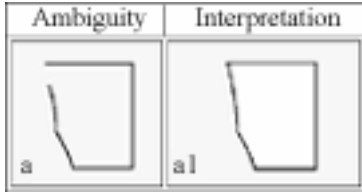


Fig. 8. Ambiguity solved by repetition method using modalities

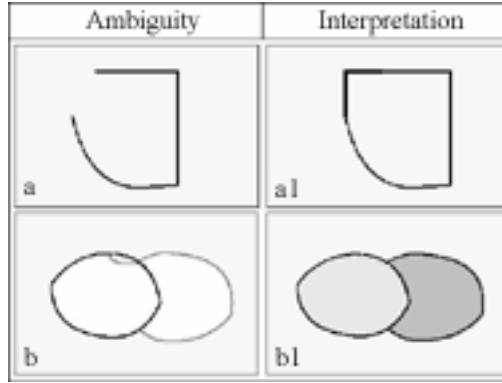


Fig. 9. Ambiguity solved by repetition method using Granularity of Repair

The second sub-method of repetition, the granularity of repair method, is described below. It can be used to solve ambiguities due to the inaccuracy of the user’s tracing. It focuses on correction of a component part of the sketched object. If, for example, the user is drawing a rectangle and he/she does not close the shape (Fig. 9a) and the system cannot provide any interpretation, then the user partially has to repeat his/her drawing completing the drawing of the boundary of the rectangle (Fig. 9a₁). In an analogue manner the inaccuracy of the user’s tracing can produce undesired polygons and polylines (Fig. 9b). The user has to partially repeat his/her action deleting the undesired polygons or polylines (Fig. 9b₁).

The second method for the a-posteriori resolution of ambiguities is the choice. This method consists of a dialogue with the user that enables the system to identify the correct interpretation of each ambiguity. The system shows the candidate interpretations to the user, which can choose the best one according to his/her intention. This method provides a feedback according to the user’s behaviours and preferences and it can be used to solve all the ambiguities introduced in the second section. For example, let us consider ambiguities due to the deleting and re-tracing actions where the user deletes and redraws the common boundary between two polygons (Fig. 10). In this case the system proposes to the user three different interpretation: i) two different polygons A’ and B’ (Fig. 10a₁); ii) a polygon and a polyline (Fig. 10a₂); iii) three polylines (Fig. 10a₃). The user can choose the best interpretation among the three previously shown cases. In sketch-based interaction using the choice method the system often proposes the beautification approach. This method can solve ambiguities due to the inaccuracy of the user’s tracing, because it beautifies the user’s tracing and shows all the possible beautified interpretations. For example, if the user draws a rectangle and he/she does not close the shape (Fig. 11 a), the system can provide to the user two interpretations: i) a beautified rectangle (Fig. 11a₁), ii) a beautified polyline (Fig. 11a₂). Finally the user can select one of these two interpretations according to his/her intention. A second example is given by the inaccuracy of the user’s tracing that produces undesired polygons and/or polylines (Fig. 11b). The system beautifies the sketch deleting the small undesired polyline.

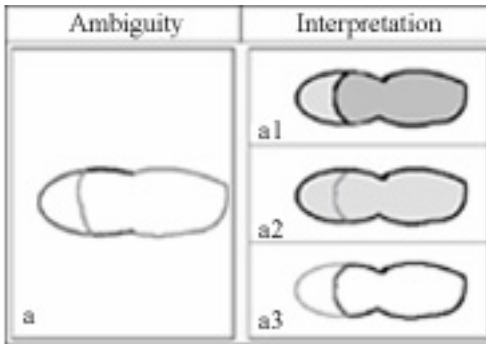


Fig. 10. Ambiguity solved by choice method

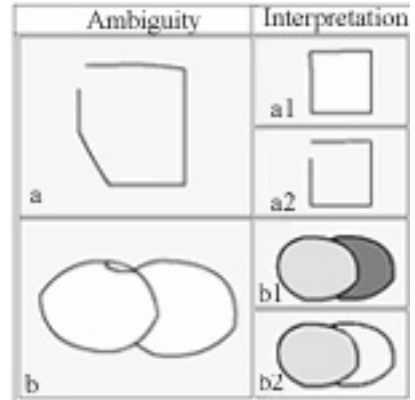


Fig. 11. Ambiguity solved by beatification method

Then the system presents to the user two interpretations: i) two polygons (Fig. 11b₁) or ii) a polygon and a polyline (Fig. 11b₂). The user selects one of these interpretations.

3.3 Approximation Resolution Methods

Ambiguities due to inaccuracy in the Human Computer Interaction behaviour can be differently solved using approximation resolution methods of ambiguities such as: i) Thresholding, ii) Historical Statistics, iii) Rules. The approximation resolution methods of ambiguity do not require any user disambiguation process.

Below the different approximation resolution methods are detailed showing their main features and giving some examples of ambiguities that can be solved.

The first method is the Thresholding [11]. It can be used to solve the ambiguities due to the inaccuracy of the user's tracing. In particular the probability of correctness of the user's input can be expressed using a probability, which can be compared to a threshold. In fact, a recogniser can return a confidence score measuring the probability that a user's input has been correctly recognized. If this confidence measure is below some pre-defined threshold the system rejects the interpretation.

For example, dealing with ambiguities due to the missing closure of a polygon (Fig. 12a), the method intercepts the missing closure as an inaccuracy and interprets the sketch as a polygon (Fig. 12a1). A second example is given by the inaccuracy of the user's tracing that produces undesired polygons and/or polylines (Fig. 12b). The system deletes the small undesired polyline and presents to the user two interpretations: i) a polygon and a polyline, and ii) two polygons (Fig. 12b1), in Fig. 12 only the last interpretation is shown.

The method of Historical Statistics can be used if the confidence score is not available or they can be wrong, probabilities can be generated by performing a statistical analysis of historical data about ambiguities. Usually historical statistics may provide a default probability of correctness for a given interpretation when a recogniser does not. This approach may use a confusion matrix, which is the matrix







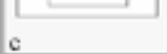

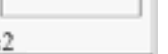
Ambiguity	Thresholding	Historical Statistics
 a	 a1	 a2
 b	 b1	 b2
 c	 c1	 c2

Fig. 12. Ambiguity solved by Thresholding and Historical statistics



Ambiguity	Interpretation
 a	 b

Fig. 13. An ambiguity solved by Rules

whose values give the estimation of the number of times that the recogniser confused the Visual Symbol. That is, if thresholding cannot disambiguate the freehand sketches, historical statistic data about correctly interpreted ambiguities can be used.

The Rule method is usually preferred when the interpretation process requires to take into account the contextual information. Freehand sketches are complex to recognize. They are often ambiguous to be interpreted without contextual information. Sometime the management of the ambiguities can require the use of the context, because thresholding and/or statistical approaches cannot be sufficient for the ambiguity solution. An example of use of rules is in [12]; it is used in speech-based interaction. Rules can be used to solve all classes of ambiguities introduced in the second section. As an example let us consider the ambiguity due to the deleting and re-tracing action (Fig. 13a). A rule could establish that if the deleted and re-traced pixels belong to the common boundary between two potential polygons then the system has to interpret the sketch as two polygons (Fig. 13a₁).

4 Conclusions

The diffusion of mobile devices for the different human and social activities is mainly due to the fact that they make communication intuitive and spontaneous. This can produce ambiguities in the interaction process and the need to manage and to solve them.

This paper proposes a classification of ambiguities in human computer interaction using a free-hand drawing approach. In particular five classes of ambiguities are considered: i) ambiguities due to crosses in a stroke; ii) ambiguities due to the over-tracing of different strokes; iii) ambiguities due to the intersection of two polygons; iv) ambiguities due to the inaccuracy of the user’s tracing; v) ambiguities due to the deleting and retracing actions. Several classes of methods to solve these ambiguities are proposed too. They are: i) prevention methods of ambiguities, ii) a-posteriori resolution methods, iii) approximation resolution methods. These methods can use one or more than one modality. A taxonomy between the proposed classes of

ambiguities and the classes of solution methods is given. Ambiguities and solution methods involving different cooperation approaches between modalities will be discussed in future works.

References

- [1] Kitawaki, N.: Perspectives on Multimedia Quality Prediction Methodologies for Advanced Mobile and IP-based Telephone. Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, Mainz, Germany, (2004) pp. 1-8.
- [2] Pham, B. Wong, O. : Handheld Devices for Applications Using Dynamic Multimedia Data. Proceedings Graphite, Singapore, (2004) pp. 123-130.
- [3] Avola, D. D'Ulizia, A. Ferri, F. Paolozzi, S. : Exploring Mobile Applications for Educational Purposes. IEEE International Workshop on multimodal and pervasive services MAPS'06 (2006) pp. 13-17.
- [4] Mahoney, J.V. Fromherz, M. P. J.: Three main concerns in sketch recognition and an approach to addressing them, AAAI Spring Symposium on Sketch Understanding, Stanford, CA, March (2002).
- [5] Ferri, F. Grifoni, P.: Interpretation and ambiguity resolution in Sketch-Based Interfaces. Proceedings of the Ninth International Conference on Distributed Multimedia Systems, DMS 2003, Miami USA (2003).
- [6] Mankoff, J. Hudson, S. E. Abowd, G. D.: Providing integrated toolkit-level support for ambiguity in recognition-based interfaces. Proceedings of ACM CHI'00 Conference on Human Factors in Computing Systems, (2000) pp. 368 -375.
- [7] Mankoff, J. Abowd, G.D. and Hudson, S.E.: Interacting with multiple alternatives generated by recognition technologies. Georgia Tech GVU Center, Tech. Rep., GIT-GVU-99-26 (1999).
- [8] Mankoff, J. and Abowd, G.: Error Correction Techniques for Handwriting, Speech, and other ambiguous or error prone systems. Georgia Tech GVU Center Technical Report, GIT-GVU-99-18 (1999).
- [9] Dey, A. K. Mankoff , J.: Designing mediation for context-aware applications. ACM Trans. Comput.-Hum. Interact. 12(1), (2005) pp. 53-80.
- [10] Martin, J.C.: Towards, Intelligent cooperation between modalities. The example of a system enabling multimodal interaction with a map, Proceeding of the IJCAI-97 workshop on Intelligent Multimodal Systems. August, 24th. Nagoya, Japan, (1997).
- [11] MacKenzie, I.S. and Chang, L.: A performance comparison of two handwriting recognisers, Interacting with Computers 11 (1999) pp. 283-297.
- [12] Baber, C., and Hone, K. S.: Modelling error recovery and repair in automatic speech recognition. International Journal of Man-Machine Studies 39, 3 (1993), pp. 495-515.

Supporting Mobile Activities in a Shared Semantic Ambient

Fabio Pittarello and Augusto Celentano

Università Ca' Foscari di Venezia, Dipartimento di Informatica
Via Torino 155, 30172 Mestre (VE), Italy
{pitt, auce}@unive.it

Abstract. We discuss an approach for modeling human activities in complex real environments, considering the delivery of services as a function of the semantic features of the environment and of the interaction between the users and their social networks. We propose an architecture supporting such a model, and discuss a case study about cooperative learning in a cultural heritage site.

1 Introduction

The execution of complex activities in wide and diversely structured environments is well supported by systems aware of the user and environment context, able to change their behavior to best fit the specific situation. Context aware adaptive systems are becoming common, and standards for describing the context are being consolidated [1,2,3]. Such systems are modeled according to two main paradigms: (1) systems that adapt the behavior of an application, designed to be adaptable, to a set of parameters that define the specific context [4,5]; and (2) systems that build an instance of an application by searching and integrating services appropriate for the specific context [6,7]. A common aspect of the two paradigms is the analysis of the “situation” and the choice of an appropriate set of actions. The choice is driven by parameters bound to the user, the devices, the network, the ambient, the location, the time, etc..

This paper discusses an approach for modeling human activities in complex real environments, considering the delivery of services as a function of the semantic features of the environment and of the interaction among the users and their social networks. The adaptation of services based on a set of independent parameters may be insufficient for activities executed in complex environments, composed of several tasks logically organized, extended in time and requiring cooperation between the users. In such cases, context adaptation must rely on the semantics of the whole application and of the ambients in which it is executed. As an example, let's examine the delivery of several types of information about an urban area. Traditionally, adaptation could specialize the search and presentation of information according to the user device, giving:

- on a conventional Internet access, full search capabilities with a multimedia presentation of usually limited quality to accommodate also for poor network performances;

- on a location specific kiosk, restricted search capabilities constrained to the ones relevant in the location, with full multimedia presentation since the multimedia material can be locally stored;
- on a PDA, information abstraction and progressive disclosure on demand, taking into account both the limitations of the device and the bandwidth and costs of the wireless network;
- on a cellular phone, limited, guided search capabilities with short vocal or text messages.

Shifting from context awareness to ambient awareness, services are delivered not only as a function of local properties of the user and her/his environment, but also according to the ambient general properties and role, i.e., according to the ambient semantics. The example above could therefore be discussed with a different perspective about the variants needed for an on-line information service. Assuming that conventional Internet access to a Web server through a personal browser can be tailored to the user context as discussed above, other ambients may offer the same service in different ways:

- in a public place, such as a mall, a station, a square, only general info of public interest is delivered through large displays and unattended kiosks;
- ambients characterized by a specialization of their functions deliver only information related to the function. For example, in a theatre atrium information about the theatre program is displayed, while in a museum information pertinent to the museum is given;
- in some ambients (e.g., a church) the specificity of the function and the mood of the place can suggest to inhibit at all information delivery;
- finally, the modality of delivery, e.g., audio messages or video displays, can also be bound to the average user attention in the ambient, for example by delivering audio messages in a station and video messages in a mall.

An additional level of adaptation comes from the knowledge of the user social environment, which may extend the information available to the user and his/her capability to perform a task, setting up a cooperation among the user and his/her network of contacts. Considering the user social network improves the set of available services, providing additional information by integrating the communication in the user social network with the results of the local services.

2 Modeling Activities in Semantic Spaces

We make some assumptions, without entering into detailed specifications:

- the physical ambient in which user activities take place is part of an urban environment, whose topological and geometric features are known, and can be classified according to a suitable ontology [8];
- the users are primarily interested to navigating the ambient for discovering (or receiving on explicit query) information, possibly in multimedia format, through multimodal interaction;

- the ambients in which users navigate have distinguished roles in terms of purpose and social relations. For example, they can be public or private ambients (e.g., a square vs. an apartment) according to some access control policy, or can be connection or action spaces (e.g., in a museum, an aisle vs. an exhibition room), according to the absence or presence of specific activities other than movement.
- the user activities are not isolated, but are elements of a plan (which can be completely defined or dynamically built) mixing activities and navigation in the environment, and requiring the execution of a number of different tasks [9];
- finally, the users themselves are not isolated, but immersed in a network of social relations which allow (and may in some cases require) interaction among the participants in order to execute the activities.

The semantic description of the environment, possibly mapped to its physical description (e.g., through a corresponding 3D geometry) is necessary for presenting the users a multimodal description of the areas that compose the physical environment, and of the appliances the user can interact with. Cues for moving in the environment and interacting with it, given according to a hierarchy of environment locations, are more effective if the user understands the relations between the environment topology and the meaning of its components, such as rooms, places, stairs, etc. [10,11].

The semantic description can also answer specific information needs: e.g., in a cultural tourism application a user, during a guided tour, might be interested in the architectural details of a building, or in the logical organization of the work of arts contained in a suite of exhibition rooms.

Cooperation among users can be exploited in two ways: locally and remotely. Local cooperation takes place in the environment and its surroundings. Several users may be involved in the same activity, or can exchange mutual help to improve a task execution. The network of social relations of a user in practice extends its environment with parts of the environments of the cooperating users.

3 Applications, Activities and Services

Generally speaking, we can define an application as a set of cooperating services executing tasks and exchanging data, providing computational and information support to human activities in a coordinated plan of actions. We do not assume any specific service architecture, such as Web services, but remain at an abstract level. Services execute (sub-)tasks, receive and deliver information and objects, process information content and presentation, etc. according to some plan. Services may also be executed by humans: for example, an information delivery service can be offered, in different contexts, by an automatic answering system, an interactive kiosk or a human guide.

Different types of services are offered to users populating an ambient: field services and local services are characterized by being, respectively, accessible in a wide area, possibly by more users at the same time, or at a specific location,

usually by one user at a time. Individual and social services are characterized by being, respectively, executed by a single user or cooperatively by a set of users.

Field services. Field services are associated to a part of the ambient geometrically defined as an area, e.g., a street, a square, a room. They are accessible, in principle, from every location inside the area. They are not bound to physical contact with the user, and do not perform transmission of concrete objects, such as tickets, money and goods. Examples of field services are:

- information delivery services in the waiting room of a railway station through large displays, visible from every room point;
- audio messaging services through loudspeakers unconditionally directed to the public in the ambient;
- services delivered through wireless communication technologies such as Wi-Fi of mobile telephony, possibly contextualised on the user location.

In general, the service dealer location and the user location inside the field is not relevant for the service execution, even if different locations can affect the quality and the performance of the service.

Local services. Local services are associated to an appliance of the environment, a physical or virtual artifact which contains the interface between the user and the service. They are accessible only in proximity of the appliance, which acts as the service dealer, may be bound to a physical contact between the user and the appliance, and may exchange concrete objects, such as tickets, money, goods. Examples of local services are:

- ticketing services delivering concrete tickets (as opposed to ticket reservation services), possibly as the final step of a sequence of reservation-issuing-delivering services;
- ATM services, and in general services related to physical exchange of money or goods;
- access control services, which are executed at specific site entrances;
- services based on proximity identification, such as Bluetooth or RFID based services, active close to an object or a location.

Individual services. Individual services are activated by a user as part of an application requiring neither cooperation nor sharing with other users. Local services are often individual services, and the effect of their execution is usually limited to the user that has executed the service. A typical example is an ATM service.

Social services. Social services are executed by a user in cooperation with other users, which can be located locally or remotely. Cooperative services are often field services, due to the need of sharing the service access from the different places where the users are. They could also be executed as the result of a local service, whose completion requires (as a subtask) the access to a wider community, through remote communication tools. An example of social service is the search for information needed to complete the execution of a task, which may require the intervention of other users expert of the task.

4 A Case Study: Cooperative Learning in a Cultural Heritage Site

The following example illustrates how the integrated knowledge of an ambient, services and social network can be used in the context of a cooperative learning experience. The case study considered is that of a class of students visiting a cultural heritage site for an on-the-field learning experience. Students are divided into small groups and the task of each group is to compose an assignment related to a specific topic, e.g., expressionist paintings. Each group is provided with a mobile device as a support for completing the task.

The knowledge of the physical features of the ambient and of the information associated to its components constitutes a first layer that can be used by an implementation architecture for helping the students to find and reach specific locations compliant with the goal of the initial task [11]. Once in the correct location, the student can access information related to the surrounding artworks. Group discussion is catalyzed by this information level and leads to the production of a preliminary version of the assignment that is the result of the local discussion (i.e., in the group).

Visiting a cultural heritage site implies the access to a set of resources that can be available with limited spatial and temporal constraints, but also to resources that can be conditioned by such limitations: for example, a pavilion with a specialized installation could be accessed only by one person at a time, or a video could be displayed in a room with a very small number of seats. In such situations the knowledge of a network of services for mapping such resources, monitoring and reserving them can be useful for optimizing the time available to students for completing their task.

When the information associated to the environment has already been examined without giving the users a satisfactory level of knowledge, the access to external networks, such as communities and even the web, can be useful to identify additional local (e.g., another class of students or an expert currently visiting the same exhibition) or remote (e.g., people whose profile is compliant with the students needs and that are available for lending a hand) individuals that could give additional help.

The knowledge of the environment, and in particular of the environment the group of students are currently visiting, is used as an additional cue for finding adequate support in the social network, informing the available remote human resources about the context of the group, giving it a better support.

Such support is in fact an enhancement of the activities characterizing the online discussion groups, where the participants usually have to explain not only their problems, but also the context where they are; the context is a useful information for the remote users trying to help them. For example, a remote user knowing about the museum in which the group is working, could make a precise suggestion, such as: *“Look at room 5, next to your room, where artworks produced by the masters of the Bauhaus art school are displayed. It could be useful to compare them with the expressionist works you’re considering”*, while

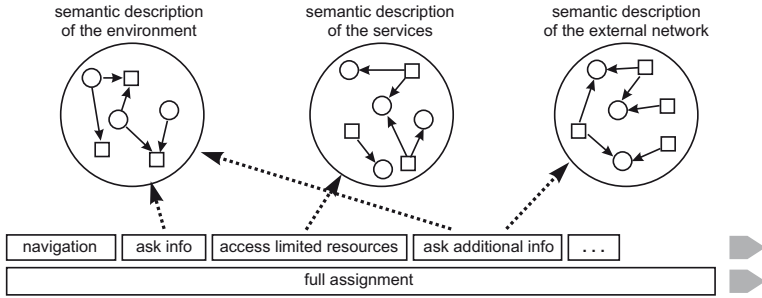


Fig. 1. Activity, ambient, service and network relations

a user unaware of the specific group context could only give a generic support independent of the museum content.

5 Relating Activities to Ambients and to Social Networks

Figure 1 shows a scheme of the task the group has been assigned to and the relations with the different semantic descriptions that are accessed in order to reach the goal. The task has been split in several subtasks, shown in the figure, that are required to complete the assignment. For the sake of simplicity Figure 1 describes a specific sequence of tasks and their relations with the semantic descriptions. While the relations between a given subtask type and a specific semantic description are fixed, the different subtasks can be iterated and can be organized in different sequences, according to the interaction patterns determined by the users' actions. Figure 1 evidences several relations:

- users moving around to find right places and artworks (*navigation* task) take advantage of a semantic description of the environment while moving; written or spoken hints are presented on the mobile device for guiding the users to the relevant locations, such as “*go straight for 3 meters, turn right and reach the hall with a circle of columns in the center*”;
- users searching for information about museum spaces and artworks (*ask info* task) require also the access to the semantic description of the environment and to the associated multimedia information, such as information associated to the semantic object *Alter Klang* by Paul Klee, belonging to the class *artwork*, or to the semantic space *Early Expressionism Room* belonging to the class *room*, for receiving a first set of information that they can use for their assignment;
- users activating the *access limited resources* task take advantage of the semantic description of services; such knowledge is used by the users for accessing the services and exchanging information about their availability and reservation (e.g., “*reserve 4 seats for the projection about Franz Marc starting at 10.00 a.m.*”);

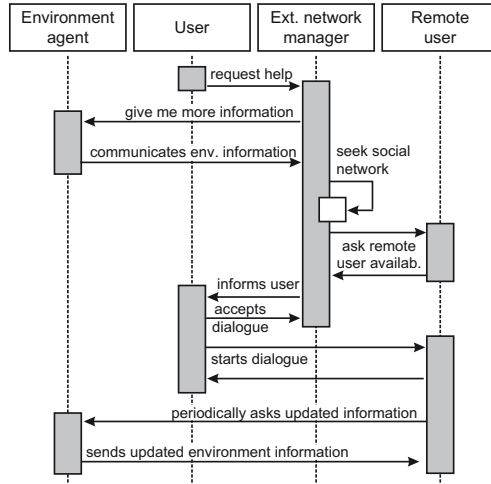


Fig. 2. Communication between a user and the external network

- users executing the *ask additional info* task can take advantage of an external network of resources in order to complete their task, using both multimedia resources available on the web (such as search engines, on-line encyclopedias, dictionaries, etc.) and human resources that are selected on the semantic representation of the social network on the basis of the users’ profiles and other parameters.

The access to the external network represents the most complex situation and includes also the access to the semantic representation of the environment for assisting the user. Figure 2 shows a sample of the communication protocol between the different subjects in the external network. While the network is usually accessed to get multimedia resources available on the web, the example is focused on the social network that represents the human component of the external help resources.

The access to the social network is triggered by a request for additional info, initiated by the group of students through their mobile device. Such request is received by a software agent, the *external network manager*, that seeks the semantic description of the social network using as input the students’ profiles, the keywords related to the topics requested by the students and the description of the environment the students are currently visiting. The environment information is asked to another agent, the *environment agent*, that sends a machine-readable description of the part of the environment the subjects are currently visiting. Such information is processed by the environment agent on the basis of the semantic description of the environment and on the specification of the current location of the users, identified through their mobile device. Such information represents an enhancement of the context associated to the students

because it includes an explicit description of the environment context that may have been omitted or only partially described in the students' requests.

The external network manager matches all such information with the semantic description of the social network, in order to find the potential helpers. Generally speaking, such human resources have declared in their profile the permission to be requested for help about certain topics by specific categories of users. They are divided in local and remote users on the basis of their location in relation to the location of the people requiring help. A different communication protocol is activated according to the helpers' location. The example of Figure 2 considers the case of a remote user (a helper far from the students, who therefore can only interact with them by sending information to their mobile device) who is contacted by the external network agent to confirm his/her current availability. The network manager, on the basis of a positive acknowledgement, notifies the remote user profile to the students that may decide to start the dialogue. The communication happens primarily between the group of students and the remote user that may freely decide the contents and the duration.

A complementary support role is offered by the environment agent that can be periodically queried (upon the students' group permission) by the remote user for sending updated information about the users' environment. Such information can be profitably used by the remote users for giving a better support, since the knowledge of the part the museum the students they are currently visiting can guide the group to interesting local resources.

6 An Architecture for Mobile Activities in Shared Ambients

Figure 3 illustrates the functional architecture of a system supporting mobile activities in shared ambients. Services are distributed in the environment as local or field services. The same service can be executed in different ways under different contexts. Also, different services can be available to users under different contexts. Therefore, the visibility of services and their access is mediated by a semantic description (*service ontology*), which manifests to the user only a set of *qualified services*, i.e., services which can be properly executed in the user and ambient context.

The ambient is described at two levels, both as a base geometry and at a semantic level, obtained by filtering the base geometry through an ambient ontology which gives meanings and roles to the objects and places of the ambient, defining also the relationships between places and objects [11]. Each place and each ambient object can also be linked to a set of multimedia/multimodal information (not shown in Figure 3 for simplicity), which is delivered by proper services through the suitable and context compliant communication channel when information services are required.

The semantic representation of the ambient is matched against the collection of qualified services, providing the set of services and the related communication channels that are compatible with the semantics of the ambient. The user receives

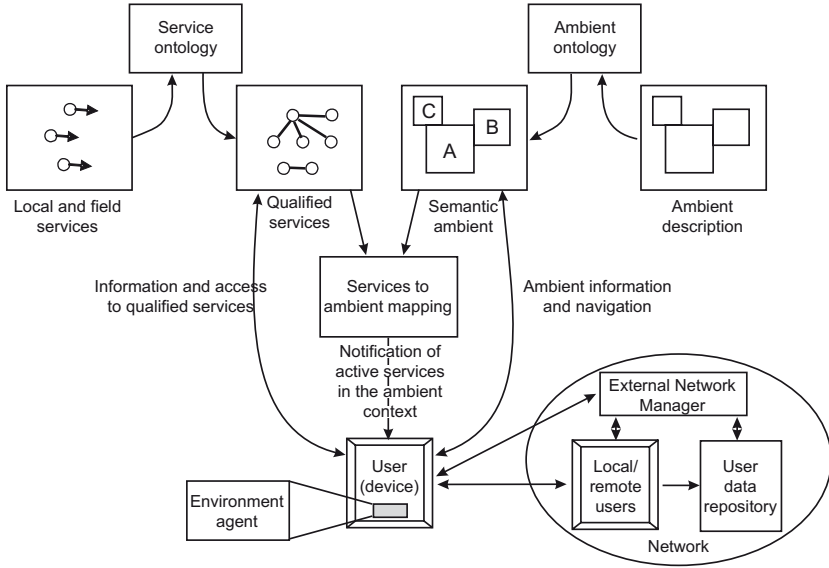


Fig. 3. An architecture for mobile services in shared ambients

on his/her device information about the available services, depending on the location, the semantic of the place, his/her own profile and history (collected by sensors and interaction patterns) and the device itself. The user can then access the available services.

The user can also access a network of external relations, where remote users may be queried for support. Access is managed by the *external network manager* that seeks the potential helpers in the user data repository, filled with data coming from people that has accepted to publish them. The manager asks to the potential helpers and to the user a confirmation of their availability for starting a dialogue; on a positive acknowledgment, a direct communication channel between the users is opened. Alternatively, if the helpers are close to the user location, the manager may notify them of such situation for a direct contact. In the case of remote communication, the channel opened by the external network manager can be used by the helpers also for accessing the *environment agent*, a software component embedded in the user device that knows information related to the environment, the qualified services and the user data (e.g. location). As explained in Section 5, such information can be useful to remote helpers for improving their assistance to the user.

7 Conclusions

In this paper we have proposed an approach to the design of mobile pervasive services based on three key elements: (1) the semantic characterization of the ambients where services can be accessed; (2) the semantic description of services

matching the services available in the environment, active at a given moment, with the constraints for their delivery according to different context parameters; (3) the opening to an external network of information and remote users to improve the number and quality of available services.

The semantic knowledge of the different ambient components is cooperatively used to reach several goals: the knowledge of the environment and of the services is used to give a first level of support to the user activities; the knowledge of the user social network is used to find additional support; finally the knowledge of the user ambient is shared to give to the components of the social network additional cues for supporting the demanding user.

Our next step will be the design of a more integrated cooperation between the available services and the social network, with the aim of adapting the services according to the information provided by the user contacts. For example, user preferences might be inferred from community profiles, and hints provided by other users in blogs and forums might be used to select and adapt the more suitable services.

References

1. Chen, G., Kotz, D.: A survey of context-aware mobile computing. Technical Report TR2000-381, Dartmouth College, Department of Computer Science (2000)
2. Dey, A.K.: Understanding and Using Context. *Personal Ubiquitous Computing* **5** (2001) 4–7
3. Held, A., Buchholz, S., Schill, A.: Modeling of context information for pervasive computing applications. In: Proc. SCI2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL (2002)
4. Lemlouna, T., Layaïda, N.: Context-aware adaptation for mobile devices. In: Proc. IEEE Int. Conf. on Mobile Data Management. (2004) 106–111
5. Villard, L., Roisin, C., Layaïda, N.: An XML-based multimedia document processing model for content adaptation. In King, P., Munson, E.V., eds.: *Digital Documents: Systems and Principles*. Volume 2023 of LNCS. Springer (2000) 104–119
6. Sycara, K., Paolucci, M., Ankolekar, A., Srinivasan, N.: Automated discovery, interaction and composition of semantic web services. *Journ. Web Sem.* **1** (2003) 27–46
7. Vukovic, M., Robinson, P.: Adaptive, planning-based, web service composition for context awareness. In: Proc. Int. Conf. on Pervasive Computing, Vienna, Austria (2004)
8. Bille, W., Troyer, O.D., Kleineremann, F., Pellens, B., Romero, R.: Using ontologies to build virtual worlds for the web. In: Proc. ICWI, IADIS International Conference WWW/Internet, Madrid, Spain (2004) 683–690
9. Pittarello, F., Fogli, D.: Modelling complex user experiences in distributed interaction environments. In: Proc. DMS, 11th Int. Conf. on Distributed Multimedia Systems, Banff, Canada (2005)
10. Mansouri, H.: Using Semantic Descriptions for Building and Querying Virtual Environments. PhD thesis, Vrije Universiteit Brussel (2005)
11. Pittarello, F., Faveri, A.D.: A semantic description of 3D environments: a proposal based on Web standards. In: Proc. Web3D 2006, Columbia, Maryland (2006) 85–95

Cultural Interface Design: Global Colors Study

Irina Kondratova and Ilia Goldfarb

National Research Council Canada Institute for Information Technology
46 Dineen Drive, Fredericton, NB, Canada E3B 9W4
{Irina.Kondratova, Ilia.Goldfarb}@nrc-cnrc.gc.ca

Abstract. This paper discusses the design of culturally appropriate web user interfaces in the age of globalization. A research study that focuses on identification and rating of visual web interface design elements that act as “cultural markers” is presented. Within this study, a web crawling technology collected data on verifiable, culture specific, web page design elements. In particular, data was collected on color usage for fifteen countries, based on the large number of country-specific websites. We found that there is a palette of certain colors that is predominantly used for website design in all countries studied. This palette was identified as an “international” color palette to distinguish it from country-specific color preferences, also found in our study. Based on these findings, international and country-specific color palettes, when incorporated in to a Web design tool, will allow designers to develop localized and international interface designs for global social networking and business applications.

Keywords: Cultural interface design, localization, usability, color palette.

1 Introduction

Globalization affects most computer-mediated communication and, in particular, user interface design for the Internet, including e-business and social computing applications. In the new global economy, as noted by Barber and Badre [1]: “As a consequence of existing international WWW users and in anticipation of potential users, usability takes on an immediate and relevant cultural context”. Nowadays users are increasingly accessing Internet applications for business, learning or pleasure using a variety of computing devices, including handheld devices, mobile phones, TV, and Internet appliances. Cultural interface design for pervasive computing is becoming an area of significant importance within the research area of human-computer interaction and visual interface design.

The importance of cultural appeal in the age of pervasive computing is growing as computing devices are becoming an essential part of user’s every day life experience, being embedded in common objects and cultural surroundings. The need for culturally appropriate interface design for Web-based e-business and e-government applications is emphasized by many researchers [2], [3], [4], [5], [6], [7]. Specifically, it is noted that the “culturability” [1], a combination of culture and usability in Web design, directly impacts on the user’s perception of credibility and trustworthiness of websites [5], [8], [9]. A culturally sensitive e-commerce framework developed by

Sudweeks and Simoff [10] lists cultural appeal as one of the four important factors impacting on sustainability of e-commerce activity, along with economic appeal, usability and general attitude towards e-commerce; thus reflecting the importance of cultural factors in e-commerce applications.

There is a growing body of evidence that supports the importance of culturally appropriate design for e-learning applications [11], [12], [13], [14], [15]. This is not surprising, considering the influence of user interface design on the usability, accessibility and acceptability of software. “Usability is the measure of the quality of user’s experience when interacting with a product or system” [16]. It includes factors such as ease of learning, efficiency of use, memorability, error frequency and severity, and subjective satisfaction. Thus, applying culturability design principles in the design process of e-learning materials is an important factor to consider.

In order to identify cultural preferences in interface design, we are conducting an ongoing study that focuses on identification of culture-specific interface design elements for a number of countries. This study is described in the next sections of the paper.

2 Cultural User Interface Study

The study investigates the usage of specific cultural markers for Website design in a number of countries, in order to incorporate the results into a cultural interface design advisor tool. Detailed description of the study is presented elsewhere [17]. For the purpose of this study, cultural markers are defined as “interface design elements and features that are prevalent, and possibly preferred, within a particular cultural group” [18]. The visual cultural markers we are investigating in this study are colors, font usage, number of images, and layout of the webpage.

This study is carried out via an automated “cultural audit” of a large number of websites from different countries. A Cultural Web Spider (Web crawler) tool, designed to extract information on culture specific Web page design elements (cultural markers) from the HTML and CSS code of websites for a particular country domain (eg: .ca for Canada, .fr for France, .jp for Japan, etc.) is used in the study [19].

2.1 Cultural Web Spidering

The Cultural Web Spider application (CWS) utilizes Google APIs Web services [19] to search for particular cultural markers on web pages of top ranked websites for a country domain. With the aid of Google Web APIs service, software developers can query more than 8 billion web pages in the Google index, directly from their own computer programs. In addition, Google API allows further restricting the search to country domain websites written in a particular language.

In this way, we were able to limit the automated “cultural audit”, of top ranked country specific domain websites, to sites written in the country’s official language (e.g. Russian, for Russia, French for France, Portuguese for Brazil, etc.) thus assuring reliability of our cultural study results. Language restriction also provides an opportunity to conduct separate audits for culture-specific websites in countries with several official languages in use, for example the cultural audit of top-ranked Canadian websites in Google index is conducted for French and English language web pages

separately. To investigate the appropriateness of our approach, and the functionality and usefulness of the cultural analysis tools we are developing, we conducted a pilot study focused on Web design color preferences for a number of countries.

2.2 Pilot Color Study

The color usage pilot study investigated design color use on the Web by studying a large number of county-specific websites for fifteen countries. The first stage of the study involved Web crawling and extraction of culture-specific information from HTML code by searching top-ranked (the most popular) pages in the Google index for a particular country and language.

A Web Crawler search was conducted for fifteen countries including Australia, Brazil, Canada (French and English), China, Finland, France, Germany, India, Italy, Japan, Russia, Saudi Arabia, Spain, United Kingdom, and United States of America. The Web Crawler was configured to extract cultural information for approximately 1000 domain names for a particular country. Search results were stored in the country database for subsequent analysis of country-specific cultural marker usage patterns.

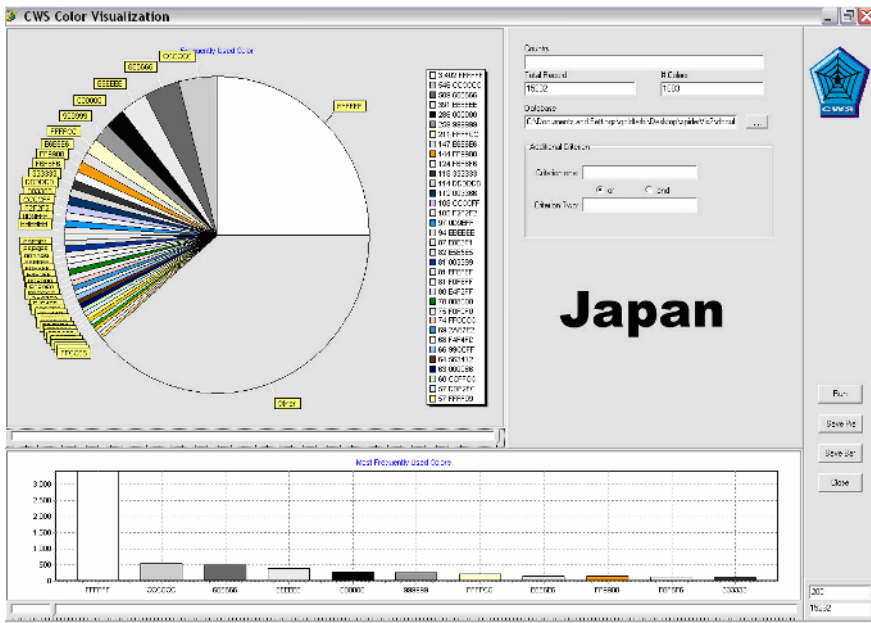


Fig. 1. CWS color visualization tool screenshot: Web page and table background colors for Japan

2.3 Visualization Tool

For the second stage of our study, the results collected by the Cultural Web Spider were statistically analyzed and visualized using a CWS visualization tool. This tool visualized the results of analysis on a particular cultural marker usage in different

countries. Figure 1 shows an example of visualization for search results on page and table background color for more than 900 top-ranked Japanese websites (.jp domain). The search was restricted to pages written in Japanese. We found that the CWS visualization tool made it easier to analyze HTML search results for color preferences and present a preferable colors palette (shown as colors in RGB format) for a particular country.

3 Color Analyzers

Web designers have at their disposal a palette of about 16.7 million colors of the HTML RGB color code to create their designs. After a testing period, we discovered that the main inconvenience of using CWS for color analysis was that the tool could not possibly show all the design colors used in a convenient and user friendly format. In order to have a meaningful presentation of a color palette, we had to limit color visualization to about 50 most frequently used colors and assign “other” label to the rest of the visualized color data, thus losing a significant portion of color usage information (Figure 1).

To resolve this issue, a color calibration tool was developed that incorporated a proprietary color classification algorithm. This tool allowed us to categorize all the colors discovered in our search into a manageable number of color categories corresponding to the user friendly “artistic” palette based on a well known “color wheel” palette.

Our “artistic” palette of 51 colors included such intuitive and easy to understand color categories as white, black, dark blue, light blue, medium blue, shaded blue, light yellow, etc. The color calibration tool functionality also allowed us to modify color categories, if needed. By using the color calibration tool, we were able to analyze the results of the Cultural Web Spider more efficiently and visualize the results via an HTML Color Analyzer. The HTML Color Analyzer represented color information we collected as a pie chart color palette for a particular country. An example of results obtained by the HTML Color Analyzer for background color usage in Japan is presented in Figure 2.

It is important to note that there are other limits imposed by the nature of the automated Web “harvesting” process. For example, it is difficult, if not impossible to automatically extract meaningful cultural information from some corporate localized websites [18]. To maintain corporate brand identity, corporate designers frequently keep the same colors and layout for all localized websites, with the only difference being images posted on these websites. In addition, for any website that has images and graphics as prominent design elements, image color information will be lost in the automated cultural analysis using an HTML analyzer, since image color information is not contained in the HTML code.

Moreover, an HTML Color Analyzer counts instances of particular color usage in the HTML code. The number of instances for a particular color does not necessarily present a true picture of color preferences, since in this case the area of color coverage is not taken into account. For example, multiple usage of a color “blue” as a cell background color in the table will result in an overall higher count of “blue” color

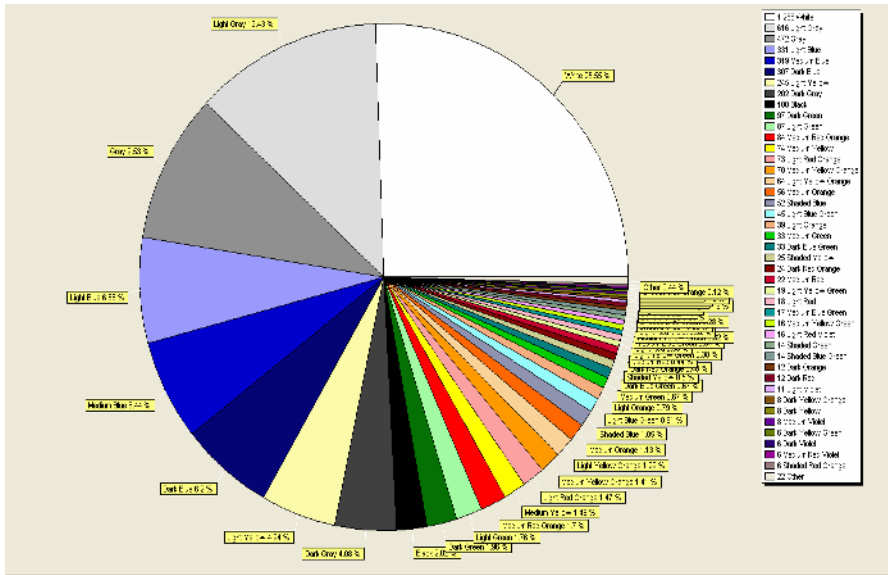


Fig. 2. HTML Color Analyzer screenshot: Web page and table background colors for Japan

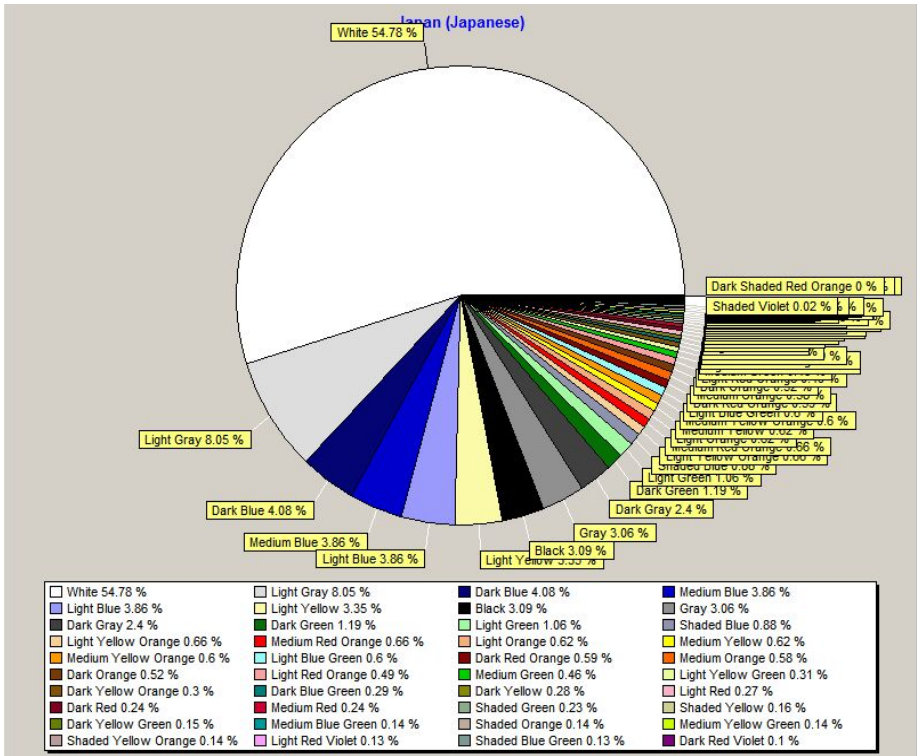


Fig. 3. Image Color Analyzer screenshot: color palette for Japan

Table 1. (continued)

Colors	Italy	Japan	Russia	Saudi Arabia	Spain	UK	USA
white							
light grey							
grey							
dark grey							
black							
shaded blue							
dark blue							
medium blue							
light blue							
light yellow							
light yellow orange							
medium yellow							
medium yellow orange							
medium orange							
medium red orange							
dark red orange							
shaded yellow							
light green							
dark green							

Color Analyzer, were examined using the following approach. We chose the sixteen most commonly used colors for a particular country, based on results obtained by the Image Color Analyzer. Thus, based on analysis of “snapshots” of Web pages, we created country-specific color palettes. After this, we cross referenced these palettes with the results obtained by using the HTML analyzer. In this way we could eliminate colors that might be present in images on Web pages, but do not correspond to color preferences we identified through the HTML analyzer. This cross-referencing process also helped us to choose a country palette of colors with both, a high coverage area and a high number of usage instances, verifying country-specific color preferences obtained by using each of the tools separately.

Results of our color usage analysis for fifteen countries are presented in Table 1. Interestingly enough, we found that the first ten colors on the list are commonly and preferentially used across all countries studied. These colors include white, black, different shades of grey, different shades of blue and a light yellow color. This color palette was named the “international colors palette”. We believe that colors from this palette could be used by designers to develop “international” user interfaces by choosing design colors that will be appropriate for a multitude of cultures. This approach would be applicable in designing Web-based e-learning applications for a broad audience of international learners. When localization is required, other,

country-specific colors could be added to the “international” palette to design an interface that will be attractive and culturally appropriate for the local audience.

For the sake of clarity, in our preliminary analysis we limited the number of colors we identified as preferable for a particular country. It is planned to expand the palette through further data analysis.

It is important to note, that in our color study we used a proprietary calibration schema that allowed us to classify instances of individual RGB colors into broad color categories such as “dark green”, “light green”, etc. Thus, individual color categories we found to be preferable for a particular country could be further expanded into a variety of RGB design colors. For example, the “dark green” color category for Japan would include many individual instances of a certain RGB color usage (such as instances of #008000 and #006600 color usage we identified in our study for Japanese websites) providing an expanded color palette that offers more freedom of creative expression for designers, within the general framework of country-specific color palette. The country-specific color palettes could be expanded based on a comparison with culture-specific color palettes and preferable color combinations that are based on historical data [20], [21].

5 Conclusions

Research shows that, in the global software development market, only careful consideration of local users’ needs will achieve long lasting success and client satisfaction with the cultural “look and feel” of the final product. This, in turn, impacts on the success of global e-business and e-learning enterprises.

However, there is a lack of software tools that can assist Web developers in creating a first draft of a cultural user interface, for a particular locale, that is verifiably culturally appropriate. In view of this, we are currently working on developing a cultural “look and feel” prototyping tool that will be based on the results of our research study, that utilizes semi-automated search and analysis of cultural markers for a large number of websites for particular locales. This study will result in the databank of cultural information, including preferred color palette information, forming the basis of country-specific Web design elements for the cultural interface design prototyping tool.

Our pilot study confirmed the feasibility of using software tools for quantitative and qualitative research on the cultural “look and feel” on the Internet. An additional outcome of this research study is that we developed a suite of tools that could be used by researchers for conducting ethnographic and cultural studies on the Internet and by marketing intelligence companies to identify cultural trends for advertising and marketing purposes.

Acknowledgements

The authors would like to acknowledge the support provided for the project by the National Research Council Canada. The authors would also like to acknowledge valuable collaboration from Dr. Roger Gervais and Luc Fournier of Centre international de développement de l’inforoute en français (CIDIF) on the Cultural Web Spider application development.

References

1. Barber W., Badre, A. N.: *Culturability: The Merging Of Culture And Usability*. Proceedings 4th Conference on Human Factors and the Web. Baskin, Ridge New Jersey (1998).
2. Hornby, G., Goulding P., Poon, S.: *Perceptions Of Export Barriers And Cultural Issues: The SME E-Commerce Experience*. *Journal of Electronic Commerce Research*, 3 (4) (2002) 213-226
3. Sun, H.: *Building A Culturally-Competent Corporate Web Site: An Exploratory Study Of Cultural Markers In Multilingual Web Design*. Proceedings 19th annual international conference on computer documentation. ACM Press New York (2001) 95-102
4. Del Galdo, E. M., Nielsen, J.: *International User Interfaces*. John Wiley & Sons, New York New York (1996)
5. Marcus, A., Gould, E.W.: *Cultural Dimensions And Global Web User Interface Design: What? So What? Now What?* Proceedings 16th Conference on Human Factors and the Web. Austin Texas (2000)
6. Becker, S. A.: *An Exploratory Study on Web Usability and the Internationalization of US E-Businesses*. *Journal of Electronic Commerce Research* 3 (4) (2002) 265-278
7. Smith, A., Duncley, L., French, T., Minocha S., Chang, Y.: *A Process Model For Developing Usable Cross-Cultural Websites*. *Interacting With Computers* 16 (1) (2004) 69-91
8. Fogg, B. J.: *Persuasive technology*. Morgan Kaufmann Publishers (2002)
9. Jarvenpaa, S. L., Tractinsky, N., Saarinen L., Vitale, M.: *Consumer Trust In An Internet Store: A Cross-Cultural Validation*. *Journal of Computer Mediated Communication* 5 (2) (1999) <http://www.ascusc.org/jcmc/vol5/issue2/jarvenpaa.html>
10. Sudweeks, F., Simoff, S.: *Culturally Commercial: A Cultural E-Commerce Framework*. Proceedings OZCHI 2001. Fremantle Western Australia (2001) 148-153
11. McLoughlin, C.: *Culturally Inclusive Learning on the Web*, Proceedings Teaching and Learning Forum 99 (1999) <http://lsn.curtin.edu.au/tlf/tlf1999/mcloughlin.html>
12. Priutt-Mentle, D.: *Cultural Dimensions of Multimedia Design for Instruction*, Proceedings National Educational Computing Conference. Seattle USA (2003)
13. Barron, A. E., Rickerman, C.: *Going Global. Designing E-Learning for an International Audience*. Proceedings ASTD TechKnowledge® 2003 conference (2003) http://www1.astd.org/tk03/session_handouts/
14. Pfremmer, R.: *Content Design Considerations for Localizing E-learning Projects*. *MultiLingual computing* (2004) <http://www.multilingual.com>
15. Seufert, S.: *Cultural Perspectives*. In: Adelsgerger, H. H.; Collis, B., Pawlowski, J. M. (eds.): *Handbook of Information Technologies for Education and Training*. Springer Berlin (2002)
16. US Department of Health and Human Services, *Usability.gov: Usability Basics*. (2004) <http://usability.gov/basics/index.html>
17. Kondratova, I., Goldfarb, I., Gervais, R., Fournier, L.: *Culturally Appropriate Web Interface Design: Web Crawler Study*. Proceedings the 8th International Conference on Computer and Advanced Technology in Education (CATE 2005). ACTA Press Anaheim/Calgary/Zurich (2005) 359-364
18. Kondratova, I., Goldfarb, I.: *Cultural Visual Interface Design*. Proceedings EDMedia 2005 - World Conference on Educational Multimedia, Hypermedia and Telecommunications. Montreal, Canada (2005) 1255-1262
19. Google: *Google Web APIs*. (2005) <http://www.google.com/apis/>
20. Cabarga, L.: *The Designer's Guide to Global Color Combinations. 750 Color Formulas in CMYK and RGB from Around the World*. HOW Design Books, Cincinnati, Ohio (2001)
21. Kobayashi, S.: *Color Image Scale*. Kodansha International (1991)

Multimodal Interactive Systems to Manage Networked Human Work

Giuseppe Fresta¹, Andrea Marcante², Piero Mussio³,
Elisabetta Oliveri², and Marco Padula²

¹ Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", CNR
via G. Moruzzi 1, 56124 Pisa, Italy
fresta@isti.cnr.it

² ITC, CNR via Bassini 15, 20133 Milano, Italy
{marcante, oliveri, padula}@itc.cnr.it

³ DICO, Università degli Studi di Milano via Comelico 39, 20135 Milano, Italy
mussio@dico.unimi.it

Abstract. This paper proposes a holistic approach to enhance human knowledge in distributed and strictly linked contexts. Current semantic web technologies focus their attention only on the machine side of the interaction process. Our approach considers both the human and the machine working together to accomplish a task, stressing the importance of adapting data, tools and interactions to the different skills, contexts and tasks of the users. Multimodal interactive systems based on Semantic Web technologies support the sharing of knowledge among co-workers in different real networked environments. We present a scenario in the building sector, in which mobile systems are needed to reduce loosing of information on technical drawings from the building yard to the office. The paper presents a specification method and its link to semantic web technologies to satisfy the emerging requirements.

Keywords: Multimodal interactive system (MIS), Semantic Web technologies, Human-Computer Interaction, Knowledge Sharing.

1 Introduction

Semantic web is progressively achieving the goal of becoming a universal medium for information exchange by giving meaning (semantics) to the content of documents on the Web in a manner understandable by machines [2]. This achievement, jointly with the development of mobile technologies, is the necessary condition to reach the overall goal of enabling human communication to share knowledge [12]. However, it is not sufficient because the overall goal requires that humans and not only machines access, understand and properly use knowledge wherever it is, in whichever format it has been recorded in the web and whichever device is available. One important reason of this insufficiency is user diversity. Users are diverse because of different culture, skills, physical abilities or because they operate in different contexts and dispose of different tools. Hence different users represent knowledge about a same domain available in the semantic web in different ways, and use and update it according to

different strategies [4]. In particular, co-workers in different but networked environments, need to manage the same information in a way useful for the specific context in which they are operating. Semantic web technologies support the translation of data and metadata, which are in a form that can be easily processed by a digital machine, into a form which can be easily processed by a human belonging to a specific cultural (sub)community, and viceversa. This translation considers also the different interaction ways that the contexts of use and the available devices suggest (i.e., desktop PC with mouse and keyboard, PDA with stylus).

This paper discusses an approach to system design aimed at overcoming the hurdles that arise in manipulating and using knowledge because of user diversity. Systems are specified as web documents, in which data and metadata define the tools for the management, enrichment, and recording of the specific knowledge.

The paper describes the project COL (Cantiere OnLine) which is now starting with the definition of the detailed user requirements and the usability validation of the exploratory prototype of the first tools developed. These tools have been developed taking into account the preliminary consideration and specifications collected with all the stakeholders involved.

The paper is organized as follows: section 2 outlines a view on collaboration in networked contexts through Multimodal Interactive Systems and, then, a real scenario in building sector is described. Section 3 specifies Multimodal Interactive Systems as web documents and virtual entities. Section 4 illustrates the architecture of BANCO, a MIS prototype. In section 5 we presented some related work and finally the conclusion in section 6.

2 A View on Collaboration in Networked Real Contexts Through the Web

We are developing multimodal interactive systems (MISs), which exploit semantic web technologies for enabling users to collaborate in networked real contexts through the web. Our MISs permit single users to have a tailored view of data and customized tools for managing these data: therefore users of different culture and skills can collaborate visualizing information according to their specific views of the problem at hand and to the available devices (mobile and desktop).

2.1 BANCO, a Multimodal Interactive System Supporting User Collaboration

BANCO (Browsing Adaptive Network for Changing user Operativity) [4] is a MIS which supports user collaboration allowing the exchange of documents and annotations on documents through the web. BANCO exploits the metaphor of the craft room. In a BANCO environment, a user can find all and only the virtual tools s/he needs to perform a specific activity. A user can organize virtual workbenches by selecting data and virtual tools to perform a specific task. A virtual workbench materializes as multimedia documents the data to be operated, along with the tools necessary to work on them. These multimedia and multimodal documents as well as the tools are organized and represented according to the user's habits and culture, the specific task and context in which the user operates.

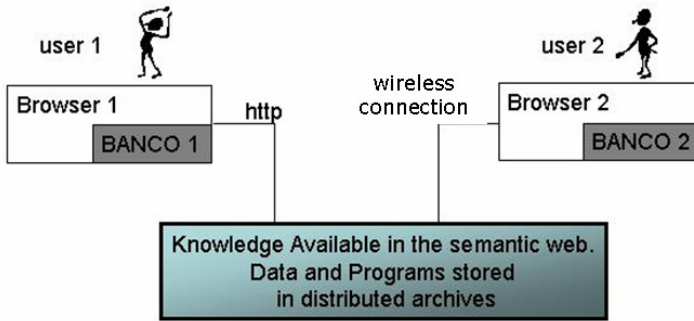


Fig. 1. Each user accesses, sees and navigates web knowledge according to his/her culture, skills and physical abilities

Figure 1 illustrates a generic scenario of collaboration through the web. User 1 accesses his BANCO 1 through his/her browser. Programs and data to materialize BANCO and manage documents and annotations are stored in a distributed and shared archive. User 1 can interact with the documents s/he is interested in and add annotations on them by using tools that are tailored to his/her culture, skills, and current task. Finally, s/he can save documents and annotations into the archive. To complete his/her task, user 1 can ask user 2 for a consultation. To this end, user 2 accesses his/her BANCO 2, tailored to his/her needs, and access documents and annotations modified by user 1. User 2 thus performs the required activity to reply to user 1 and saves the results into the shared archive. BANCO networked system exploits the idea of wiki: it permits to users to ease write and editing collectively a multimodal and multimedia document. Differing from usual wiki, users can perform also graphical editing on documents such as drawings.

In the following section we contextualize the generic scenario described above into a specific one, in which different stakeholders collaborate through the web exchanging information on the variations on the field (the building yard).

2.2 A Scenario in the Building Sector

In the building sector different players produce information which must be shared and updated. Some problems arise in managing and updating technical drawings and orders for the supplying stores. In the offices, documents are in electronic format, on the yard, they are paper based: technical paper-based drawings of a building, are often updated in the yard but about 70% of these updating are not reported to the office. To overcome this situation, mobile tools are required [5], to permit also to workers on the yard to manage electronic documents and to send back electronic updating to the office. A typical scenario (Figure 2), considers two main locations: the head office of the enterprise or the office of the storehouse, and the building yard. In the office an operator interacts with a data archive related to the building (technical drawings and

documents) through a desktop PC, which represents the data on a large-sized display. On the yard, the foreman operates with an interactive environment on a mobile device in which the same data related to the building are represented on a small display. Both the operator in office and on yard, report their activities by annotating the documents on the screen.

Figure 2 illustrates the whole scenario, where electronic documents are technical drawings of buildings, temporary annotations and annotated technical drawings. The foreman loads on the mobile devices a reduced version of electronic documents s/he needs: i.e. the technical drawings of the floor of a building (step 1). In the yard, the foreman annotates building variations on these drawings through her/his PDA (step 2). These drawings will be small-sized to allow taking the annotations and upload them via wireless connection or via wired connection to a desktop PC (step 3). In both cases, the annotations are sent to and stored into a temporary archive. We are now designing a subsystem for specific semantic management of the annotations. The last step of the scenario concerns an operator on a desktop PC in the office which evaluates the recorded annotation and updates the original electronic documents accordingly; s/he therefore saves the final electronic documents in an archive of the annotated documents (step 4).

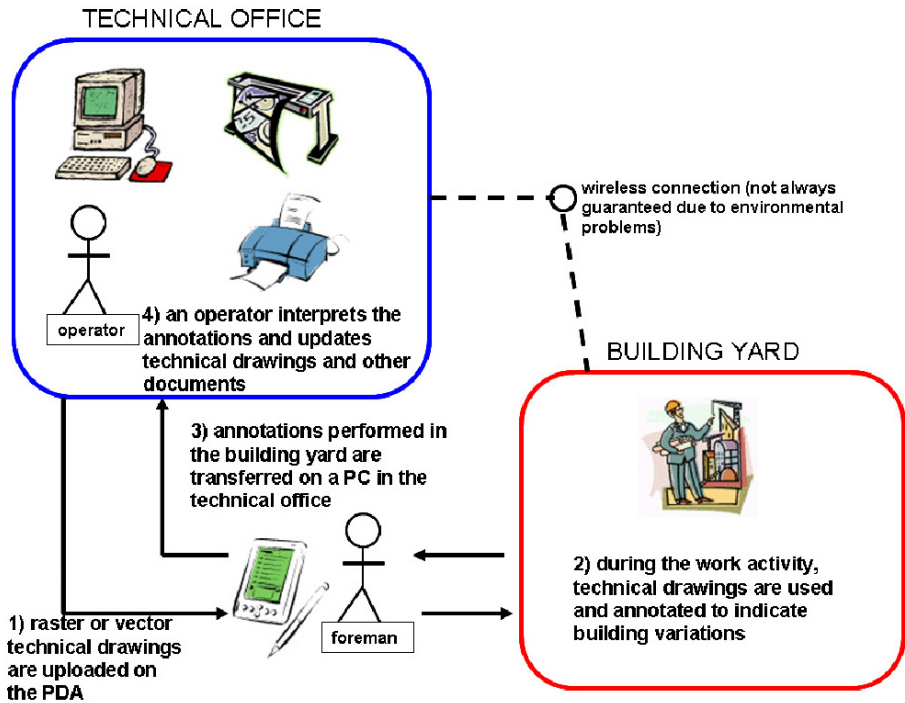


Fig. 2. The building sector scenario

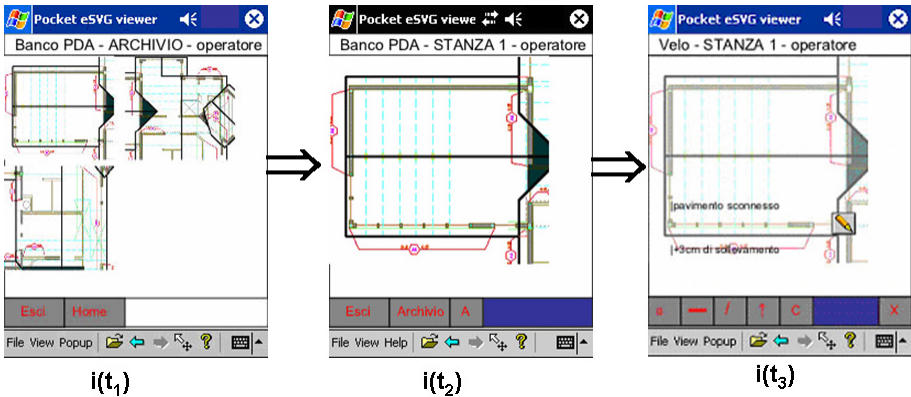


Fig. 3. Interaction sequence of the user activity

Figure 3 described the sequence of activities the foreman performs on her/his BANCO on a PDA to annotate the drawing s/he is interested in. At time t_1 , the foreman accesses the archive on her/his PDA and selects with the stylus the map of the room s/he is interested in (image $i(t_1)$). The system reacts by presenting the map of ROOM 1 ("STANZA 1") on the virtual workbench ($i(t_2)$). Selecting A button, foreman activates the annotation tools. A transparent layer is loaded over the room and an annotation toolbar is displayed on the bottom of the workbench: therefore, the foreman can write an annotation selecting the first button on the left on the toolbar or draw a line on the map selecting the tool for hand drawing (the second button on the toolbar). Then s/he clicks on the identified area to associate an annotation with it. The workshop reacts by showing a visual link (the pencil icon) on the identified area and allowing the user to add annotations ($i(t_3)$). Finally, the foreman saves the annotation in a local data repository if a wireless connection is not available, or in the data repository shared with workers in the office. If wireless connection is not available, the annotation will be later downloaded in the data repository accessible by the operator in the office.

The operator in the office operates on an annotated digital sketch by opening her/his BANCO system and by selecting the map annotated by the foreman from the menu on the right. The system reacts displaying the map on the screen (Figure 4). In the BANCO system for office the voice modality is enabled ("voice enabled" button on the bottom in the right side): this modality tell the user how many annotations are on the map and, when the user click on the visual link, the annotation content is read. In this way, user in office can immediately know how many changes have been performed on the map. If s/he needs to see a single annotation, s/he have to select the visual link and open the annotation manager. Through the voice modality, s/he can require to the system to read all the annotations without the need to select all the visual links: s/he can recognize and select the annotations s/he is interested in and update only the interested changes.

In this step of the project development, we are fitting the implementation to the requirements of the users which are accustomed, on the yard to writing signs and notes, while in the office to adopt multimodal interactions.

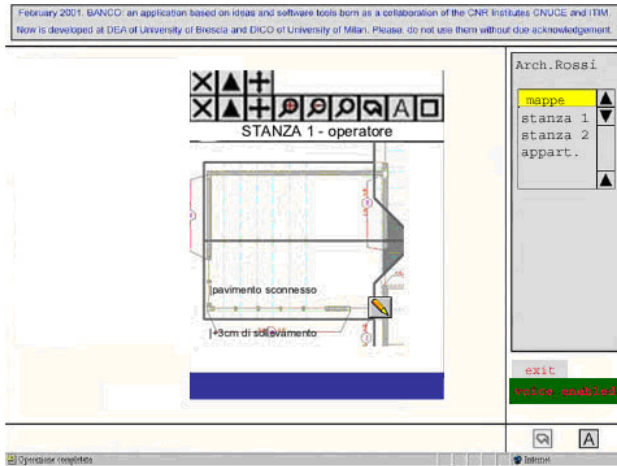


Fig. 4. BANCO for the office operator

2.3 Requirements for Improving Users Collaboration in Networked Real Environments Through the Semantic Web

From the activity described in the scenario some requirements emerge to be satisfied by a MIS supporting users to collaborate in networked real environments through the web. The system must be *interoperable*: the users can have different devices (i.e., mobile or desktop PC) that can have different configurations (i.e. different operating systems or different browsers) to access the system. The system must also be *platform independent*; for this reason, we develop systems accessible by every browser (in the prototype, the only constraint is to use an SVG compliant browser). The system must guarantee *adaptation* to user needs and, in the example, to user contexts and devices. Adaptation determines, as a consequence, other requirements. First, a balancing of the client/server traffic is needed for supporting customization and tailoring of the system. Moreover, the set of digital signs (i.e., the icons visualized on the screen but also the speech words) must be internationalized to permit the localization of the environment on the fly. Considering these requirements, we develop systems based on technologies which support portability and adaptation; furthermore, relevance will be given to the possibility to have source code and to modify and integrate it into the system.

3 Specification of Multimodal Interactive Systems

In this section, we introduce a specification method for MISs, which supports the satisfaction of the above mentioned requirements. The link between the specification method and the semantic web technologies is also discussed.

A MIS is defined by a program P written in IMML (Interaction Multimodal Markup Language), an XML-compliant language. IMML is a markup language designed to specify MISs as virtual entities. A *virtual entity* (ve) is a virtual dynamic open and possibly non-stationary system. It is *virtual* in that it exists only as the

results of the interpretation of the program P by a computer; *dynamic* in that its behavior evolves in time; *open* in that its evolution depends on its interaction with the environment; *non-stationary* in that it changes its structure in time evolving due to the interaction with the user [4]. Examples of virtual entities in BANCO are workbenches, toolbars, menus, menu items, buttons, areas including user-defined areas such as those ones identified through the hand drawing tool.

A virtual entity can be specified from two different points of view:

- as virtual dynamic open and possibly non-stationary system;
- as a set of documents in the semantic web.

In the following, we will discuss these two kinds of specification. The first is aimed at clarifying the dynamics of the interaction, the kinds of admissible evolutions of the system, and how inputs are articulated and outputs are perceived with respect to users acting in a precise environment. The second is aimed at developing a computer program that implements the system. The mapping between the two specifications is outlined in subsection 3.2 and in section 4.

3.2 Specification of a Virtual Entity as a Set of Documents in the Semantic Web

A virtual entity is generated and maintained active by a program P . In our approach, such program is specified within the semantic web as a set of XML-compliant documents. In particular, a **ve** is specified by the following documents:

1. A document that specifies the initial *content* and *organization* of the **ve**. By “content” of a **ve** we mean the list of activities associated with the **ve**. If the **ve** is an atomic **ve**, then by “organization” we just mean the description of the **ve** type (e.g. button, button panel, area, etc.). Whenever the **ve** is composed by other **ves**, by “organization” we also mean the relations existing among the component **ves** and between the **ve** and its component **ves**.
2. A set of documents that specify the initial *physical manifestation* of the **ve**.
3. A set of documents that specify the dynamics of the **ve** as reactions to input events. These documents include the *interaction managing functions* associated with the **ve**.
4. A document that specifies how to create (a) the initial state of the **ve** and (b) how to materialize the **ve** state changes. This document includes the *instantiation functions* associated with the **ve**. These functions operate having as inputs the documents (1) and (2) to create the initial state of the **ve**. The results computed by the interaction managing functions are also taken as inputs by the instantiation functions to create non-initial **ve** states.

3.3 Multimodal Interactive Systems and BANCO as Virtual Entities

In general, a Multimodal Interactive System is itself a virtual entity. A MIS is a composed **ve** (let us call it ve_{MIS}), a system of **ves** that communicate one another and with the user. These **ves**, which are sub-systems of the MIS, can be in turn decomposed into subsystems – their components **ves**. The set of **ves**, which are subsystems of a VIS, are organized in a hierarchy, defined by a relation “subsystem

of': a **ve** is a subsystem of another **ve** if MIS is a **ve** which is not a subsystem of any other **ve**, and **ves** exist which have no subsystems, the atomic **ves**. The MIS is generated by a program P_{MIS} which is not a subprogram of any other program. The state of the whole interaction process is described by relating the state of P_{MIS} at instant n with the current materialization (on the screen and by speech), defining a *multimodal sentence*, $ms_n = \langle m_n, d_n, \langle int_n, mat_n \rangle \rangle$, where m_n is the message materialized as image or voice, int_n is a function which maps cs_s in m_n into elements of d_n and mat_n maps elements of d_n into cs_s . When the state of the interaction process is $ms_1 = \langle m_1, d_1, \langle int_1, mat_1 \rangle \rangle$, and the user generates activity a , the reaction of ve_{MIS} will result into the creation of a new ms_2 , whose message m_2 appears on the screen and whose d_2 describes the new state of the program AP_{MIS} .

4 The Architecture of a BANCO Prototype

In the current implementation, a BANCO prototype is specified by an IMML program, which is interpreted by an SVG and VoiceXML compliant browser, thus generating a complex virtual entity. SVG is an XML-based language and is the W3C Standard for Vector Graphics [11]. Actually, SVG is more than a language for vector graphics, because it also provides interaction features at pixel level. IMML programs use these features to manage the interaction between users and virtual entities as well as to link annotations to precise points in the documents. VoiceXML is the W3 standard for the text-to-speech [10]: it is supported by means of the X+V technology [9]. Figure 5 illustrates the current implementation of an IMML program showing the XML-compliant documents that specify BANCO and their relations.

Let us analyze these documents:

- *S&B X+V* is the starter and bridge X+V: it is a X+V document specifying the documents to be interpreted by the VoiceXML-compliant browser to generate the vocal materialization of a particular ve_{MIS} . The S&B X+V contains a link to start the SVG starter;
- *SVG Starter* is an SVG document specifying the documents to be interpreted by the SVG-compliant browser to generate the visual materialization of a particular ve_{MIS} ;
- *DocIMML* is an IMML document specifying the static part of content and organization of the ve_{MIS} initial state;
- *DbIMML* is a set of IMML documents, each one specifying a type of virtual entity that can be instantiated during the interaction process to modify the state of ve_{MIS} : each **ve** type is specified by an instance prototype following an approach similar to that proposed in [6];
- *XML Customization Documents* are documents in an XML-based language which specify the physical characteristics of the virtual entities composing the ve_{MIS} : these characteristics are customized to the user skills, the contexts and the available device;
- *Template SVG* is a set of SVG documents, each one describing the physical materialization of a type of virtual entity composing the ve_{MIS} ;
- *ECMAScript* is a set of documents that (a) specify how to create the static part of ve_{MIS} and (b) specify the reactions of ve_{MIS} to user actions.

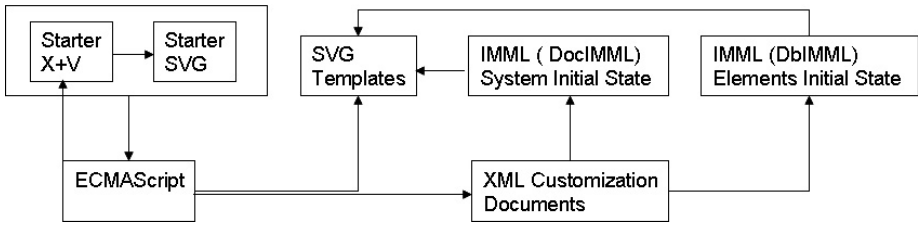


Fig. 5. The architecture of a BANCO prototype

5 Related Works

The semantic web is generally considered an evolution of the web in which the information is enriched with semantic metadata that can be processed and reasoned about by machines [2]. However, in a variety of domains, the knowledge available on the web must be not only machine-understandable, but also human-understandable, in the sense that it must be perceived, interpreted and managed correctly by users. We consider semantic web as a two-fold sources of possibilities, but also of hurdles whenever machine and human are not considered in a holistic view: in the model of Human-Computer Interaction we adopt, the PCL (Pictorial Computing Laboratory) model [3], interaction processes are determined by a cognitive system (the human) and a computing system (the computer), which in turn form a unique system, that is called “syndetic system”, i.e. a system composed by binding sub-systems of a different nature.

Many XML-based proposals have been published (TERESA [7], UsiXML [8], UIML[1]), which aim at separating an abstract description of the system interface from its concrete description. We consider a description of content and organization of the *ve* separated from the characteristics specific of a determined physical manifestation: IMML documents describe content and organization, while other documents (XML customization documents) specify the characteristics depending on user culture, skills and devices and templates determine the final materialization of the *ve* (in the proposed BANCO prototype, SVG for the visual materialization, VoiceXML for the vocal materialization).

6 Conclusions

A holistic approach is proposed which stresses the importance of the human side in the performance of human activities supported from virtual systems in networked real environments. The approach is bottom-up: it starts from the study of the user activity in the work context to derive the tools that can support and enhance the user capabilities in sharing and updating knowledge needed to perform their tasks. It also takes into account user diversity, due to different culture, skills, physical abilities, work context and available devices. The prototype here proposed is now under further development in cooperation with the users in the building sector. An authoring tool is also being developed for allowing HCI experts together with users representatives to specify their BANCO system.

Acknowledgments. The authors wish to thank the members of the COL (Cantiere-On-Line) project for their collaboration to the definition of the scenario and to the feasibility evaluation.

References

1. Abrams, M., Phanouriou, C., Batongbacal, A., Williams, S., Shuster, J.: UIML: An Appliance-Independent XML User Interface Language. Proceedings of the 8th WWW conference (1994)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* (2001)
3. Bottoni, P., Costabile, M.F., Mussio, P., Specification and Dialog Control of Visual Interaction, *ACM TOPLAS*, 21(6) (1999) 1077-1136.
4. Carrara, P., Fogli, D., Fresta, G., Mussio, P.: Toward overcoming culture, skill and situation hurdles in human-computer interaction. *Int. J. Universal Access in the Information Society*, 1(4), (2002), 288-304.
5. Fogli D., Marcante A., Mussio P., Oliveri E., Padula M., Scaioli R.: Building yard on line: a distributed and mobile system for supporting building workers. *Proc. 14th WETICE'05*, Linköping, Sweden, (2005).
6. Myers B. A., Giuse D. A., Dannenberg R. B., Zanden B. V., Kosbie D. S., Pervin E., Mickish A., Marcha P.: Garnet: Comprehensive support for graphical, highly-interactive user interfaces, *IEEE Computer*, 23(11) (1990)
7. Mori, G. Paternò, F. Santoro C.: Design and Development of Multi-Device User Interfaces through Multiple Logical Descriptions, *IEEE Transactions on Software Engineering*, 30 (8) (2004) 507-520.
8. Stanculescu A., Limbourg Q., Vanderdonckt J., Michette B., Montero F.: A Transformational Approach for Multimodal Web User Interfaces based on USIXML. *Proceedings ICMI 2005*, 259-266, ACM Press.
9. XHTML+Voice Profile 1.2, W3C, <http://www.voicexml.org/specs/multimodal/x+v/12/>, 16 March 2004.
10. Voice eXtensible Markup Language Version 2.0, W3C Recommendation, <http://www.w3.org/TR/voicexml20/>, 16 March 2004
11. W3C, Scalable Vector Graphics (SVG), Available at <http://www.w3.org/Graphics/SVG/>.
12. W3C, Semantic Web, Available at <http://www.w3.org/2001/sw/>

SIAPAS: A Case Study on the Use of a GPS-Based Parking System

Gonzalo Mendez¹, Pilar Herrero², and Ramon Valladares²

¹ Facultad de Informatica - Universidad Complutense de Madrid
C/ Prof. Jose Garcia Santesmases s/n, 28040 Madrid, Spain
gmendez@fdi.ucm.es

² Facultad de Informatica - Universidad Politecnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte (Madrid), Spain
pherrero@fi.upm.es

Abstract. GPS-based applications have become very popular during the last years, specially among drivers, who use them to find the best way to their destination. However, their use is still far from taking advantage of the wide range of possibilities that GPS offers. The SIAPAS application goes one step further by adding new functionality to the typical GPS-based map. SIAPAS runs on a PDA and it allows drivers to find a parking space that suits their needs inside a parking lot. This paper describes how the system has been designed and implemented, and shows the results of some experiments that have been carried out to test its utility and usability.

1 Introduction

Finding a parking space is a common challenge faced by millions of citizens every day. Let's imagine a driver who arrives to a shopping center looking for the place to park his car. Let's also imagine that the shopping center is on sale and therefore it is bursting with people. If the user needs to buy something quickly, something that he forgot the previous day when he did his weekly shopping, and he is also in a hurry because he just quit from his job for a few minutes, he would need extra help to find the best parking-position. The driver is not concerned with the shopping center entrances that are far away from his current location, rather he wants to choose one from several entrances near his current location and, if possible, closer to the requested shop.

A location-based application could help to this user with this problem as it would guide him depending on his current location. A crucial part of this location-based application is locating users' current location. Global Positioning System (GPS) is a widely used technology for this purpose and it is constantly being improved. With the advances in GPS and wireless communications technology and the growing popularity of mobile devices, such as PDA, the need for location-based applications has gained significant attentions.

In the last few years some similar projects have been developed in many different places with many different purposes. In fact, an overview of ad-hoc routing

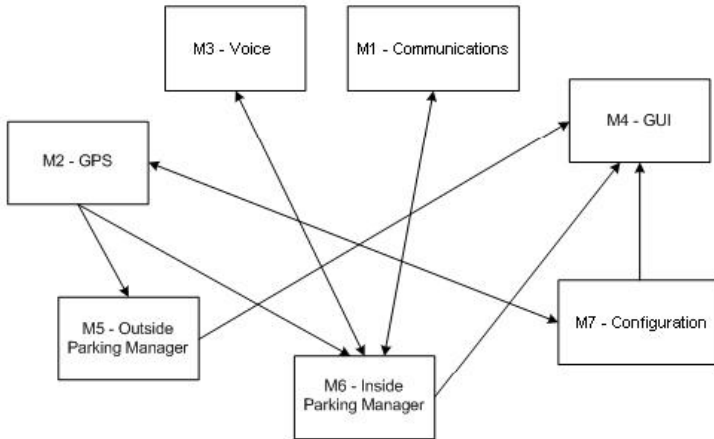


Fig. 1. Module-level system architecture

protocols that make forwarding decisions based on the geographical position of a packet's destination is presented in [1].

One of these projects is the one developed as a part of the digital campus project at the University of California, NAPA (Nearest Available Parking lot Application). This application, which finds an empty space from multiple parking lots scattered around a campus or some area like a city or an airport, is intended to reduce the bottleneck at the campus entrance, which is often a time consuming process in itself [2]. Another one is PMNET, a multi-hop wireless parking meter network that allows users to locate and navigate to an available parking space by equipping existing parking meters with wireless radio frequency (RF) transceivers and auxiliary hardware and software [3]. However, although both projects are similar in intentions, SIAPAS goes one step forward as it offers not just a location for the driver's car, but the best one: closer to the closest entrance of the shopping center.

In this paper, we will show how the system has been designed (Section 2) and we will give a few details about the implementation (Section 3). Then, we will describe how the system has been tested and evaluated (Section 4) and we will end with the conclusions we have obtained (Section 5).

2 System Design

The SIAPAS system has been designed as a set of independent modules that communicate with each other through the use of web services (see Fig. 1).

- M1 - Communications: this module keeps track of the state of the parking spaces.
- M2 - GPS: it keeps track of the car's current position, minimizing the GPS position error.

- M3 - Voice: speech-based driver's assistance.
- M4 - GUI: it manages user interaction.
- M5 - Outside Parking Manager: this module controls the global parking state.
- M6 - Inside Parking Manager: it keeps track of the parking state: parking spaces, routes, entrance and the like.
- M7 - Configuration: it manages the GPS device configuration.

This division in modules is based not only on functionality reasons for each device, but also in the global functionality. Thus, there may be parts of the same module running in different devices.

2.1 M1-Communications

This module consists of two parts (a client and a server) that communicate with each other through a WiFi network using SOAP.

The server side is based on an agent who is in charge of managing the clients' petitions to block and release parking spaces, keeping the ontology that is used to represent the parking state up to date. The clients can block the parking space they want to use so that no other driver can use it. To avoid the parking spaces being blocked without a car occupying them for too long, the agent is also in charge of releasing the ones that have been blocked for more than a predefined time (currently 30 minutes).

The client runs in the drivers's PDA, and it starts working as soon as the GPS detects that the car is inside the parking lot. It first requests the parking state, and then it blocks the parking space to be used to park the car.

2.2 M2-GPS

This module is in charge of receiving data from the GPS system and preparing them to be used by the rest of the SIAPAS application. It is structured as a conventional compiler, with a component to read data and detect lexical errors, another one to parse the sentences and a third one to obtain position and precision data and prepare them to be used by other parts of the application, usually to update the state of close parking lots or to update the drivers' position inside the parking.

2.3 M3-Voice

The Voice module was originally part of the Inside Parking Manager, but it was separated from it due to its complexity and different nature.

It has been divided in two submodules. The first one is in charge of analyzing the route that the driver must follow and create a list of events where some instruction must be told to the driver. These events include turning left and right and parking. The second submodule is in charge of checking, every time a GPS signal is received, whether the car is close to a mark where some instruction must be given to the driver and activate the speech synthesizer.

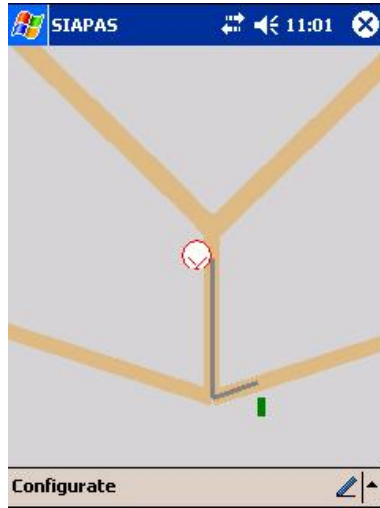


Fig. 2. Inside Parking Manager GUI

2.4 M4-GUI

This module provides a graphical user interface for three of the modules that form the SIAPAS application: Inside and Outside Parking Managers and Configuration.

For the Configuration Module, the GUI offers the possibility to change the communication port with the GPS device, the communication speed and the protocol to be used to communicate with the GPS.

The GUI of the Outside Parking Manager is the default screen that users see when they are not in a parking that is controlled by SIAPAS. In this screen the user can see which are the closest parkings, how far they are from the user and in what direction.

As for the Inside Parking Manager, the necessary data to draw the parkings are stored in an ontology in this module. These data are related to the parking itself (lanes, entrances and exits and parking spaces) and to the vehicle's position. Fig. 2 shows the aspect of this GUI, where the user can typically see: the parking lanes painted in brown; the parking spaces in green (free), red (occupied) or maroon (blocked); the vehicle, as a white circle with an arrow point inside if it is moving or a dot if it is stopped; and the route to the closest free parking space painted with a grey line.

2.5 M5-Outside Parking Manager

This module is in charge of maintaining the ontology that stores the information about the parkings that can work with the SIAPAS system. The ontology stores the name, location and size of the parking, and it is used to infer where the closest parking is or how to get there.

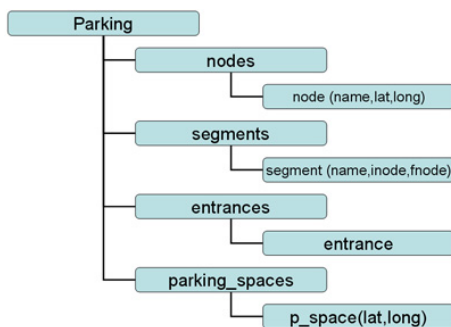


Fig. 3. Inside Parking Manager Ontology Structure

The ontology is updated every time the driver is close to a parking that is managed by the SIAPAS system. In addition, inferences are carried out after a predefined time (currently, 3 seconds) when a GPS signal is received, so the current distance or location of all parkings can be updated. The format to represent latitude and longitude is dd.mmmmmm, using a WGS84 datum. The distance is calculated using Vicenty's formula [4].

2.6 M6-Inside Parking Manager

The Inside Parking Manager is in charge of controlling what happens inside a parking, so it manages information about parking spaces, lanes and routes from one place inside the parking to another. We have used an ontology to represent this information, the structure of which can be seen in Fig. 3.

This module starts its execution when the car is close to a parking. It retrieves all the information related to the state of the parking spaces and lanes inside the parking, and it uses a slight variation of Dijkstra's algorithm [5] to figure out the most suitable parking space. Once it has been found, the Inside Parking Manager checks periodically that the driver is following the right route. If this is not the case, it calculates a new route to the parking space.

2.7 M7-Configuration

This module is in charge of managing information about the GPS hardware that is used by the system. Currently, the information that is used is the communication port (COM1 to COM6), the port speed (from 2400 bauds to 115200 bauds) and the communication protocol (only NMEA, for the moment).

When the application is launched, it reads the configuration file and tries to establish a connection with the GPS hardware. If the configuration is wrong or the PDA is using the chosen port for some other purpose, all the user will see is a message saying that there is no GPS signal available. If the user changes the configuration, the application will try to open the selected port. If it is successful, it writes the new configuration data in the configuration file; otherwise, no changes are made.

3 Implementation

The client side of the SIAPAS system runs on a Pocket PC that uses Microsoft Windows CE as Operating System. Among the different options that exist to develop software for this platform, Microsoft Visual Studio .NET 2003 has been chosen as the development environment, especially due to its good integration with the execution environment.

There are several options to implement a sockets-based communication between the clients and the server. The C++ Sockets Library has been chosen because it is object oriented and internally it makes use of POSIX libraries.

Finally, to develop the GUI there is the possibility to make use of the libraries provided by the .NET Compact Framework, a reduced set of GDI (Graphics Device Interface). Although this was the first choice, it soon became obvious that the possibilities it offers are quite low. Therefore, after analyzing different alternatives, OpenNETCF was chosen to substitute GDI, mainly because it is quite similar to GDI and easy to use (although the performance is not as good as, for example, that of GAPI).

4 Evaluation

One of the main objectives of this project is to develop an application that can be used in a short period of time, so a lot of stress has been imposed over the evaluation of SIAPAS to make sure it will be useful for the final user. The evaluation method that has been used is described in [6], and it basically consists of a theoretical validation, scenario validation and user validation.

4.1 Theoretical Validation

For the theoretical validation, the objective has been to test that the mathematical basis used in the application is accurate enough for the application to be useful. Three different experiments have been run.

Experiment 1 - Distance accuracy. Both Vicenty's formula and the Haversine formula have been used to measure distance accuracy. Vicenty's method shows that, using a WGS84 datum, longitude and latitude precision should be of 0.00005".

We have tested the algorithms running 27 tests with different data. For two points situated less than 100 meters far from each other, both methods showed a difference in measure that ranged between 1 centimeter and 51 centimeters, which can be considered a very good precision for this kind of application.

Experiment 2 - Closest entrance selection. The objective of this experiment has been to determine whether Dijkstra's algorithm always selects the closest entrance to the building or not. Fig. 4 shows the elements of the experiment, where the graph shows the parking structure, E1 and E2 are the different entrances to the building and the X shows the position of the car.

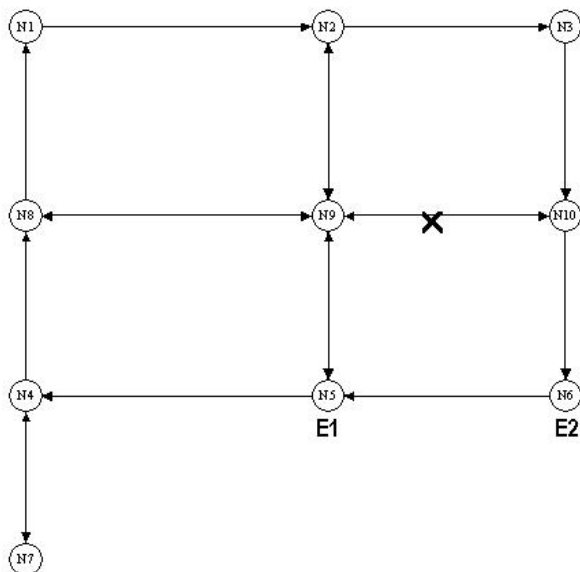


Fig. 4. Experiment 2 - Parking Structure

We have run 25 different tests, changing the position of the car, E1 and E2. In all the tests, the distance of the route selected by the algorithm was slightly shorter than the real distance travelled by the car, being the mean deviation of 5.02%.

Experiment 3 - Closest parking space selection. In this experiment, the main objective is to test that the chosen parking space is the closest one to the building and that it is the one that most people would choose. The structure of the parking is the same as in the previous experiment (see Fig. 4), and the location of the parking spaces can be seen in Fig. 5. Each one of the 25 tests that have been run in this experiment is a continuation of the corresponding test run for Experiment 2.

After running the tests, each of the solutions have been evaluated according to what a user would think of it (this has been done by the same person who ran the tests; the results of the user validation will be shown later in this paper). The values used have been: totally disagree, slightly disagree, agree, quite agree, totally agree. Then, a numerical value ranging from 1 to 5 has been given to each of the options, and the resulting average was 4.65. This average shows that, most of the times, we believe a human user would think he would have chosen the same parking space.

4.2 Scenario Validation

In this validation, the objective has been to test that the SIAPAS application offers a practical solution to the previous experiments. Fig. 6 shows the struc-

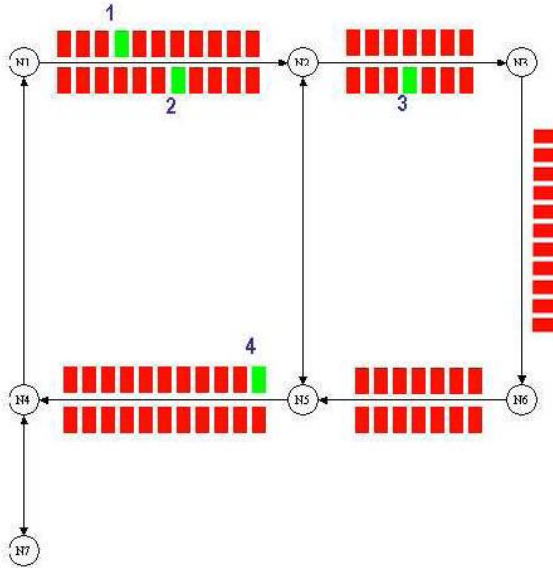


Fig. 5. Experiment 3 - Parking Spaces Location

ture of the parking, where the stars mark the entrance to the building and the rectangles show the parking spaces.

We have run 26 tests, changing the position of the car and the free parking spaces, and the results have been evaluated the same way it was done in Experiment 3. This time, the obtained average has been 4.58, which is quite close to the previous result.

After running this test, we can see that the experimental results obtained using SIAPAS are the ones expected after running the theoretical validation, and the system has shown that most of the times it is able to find the most suitable parking space for a driver. This point will be effectively evaluated in the next validation.

4.3 User Validation

This validation has been carried out using the *Performance Measure* technique [7]. Three groups of five people were made to be able to compare whether the results depended a lot on the kind of user or not:

- Users between 21 and 26 years old with low skills using hardware and software.
- Users between 24 and 25 years old with an average level experience using PCs.

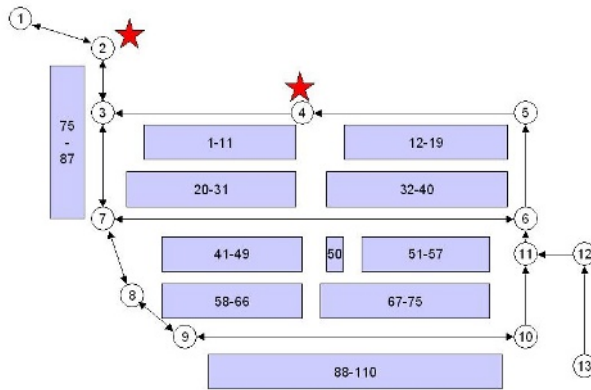


Fig. 6. Parking Structure for the Scenario Validation

- Users between 21 and 28 years old with high skills using PCs, PDAs and GPS applications.

Each user had to park his car twice, first without using SIAPAS and then using it, and we measured the time it took since they got into the car till they arrived walking to the building entrance. After that, the users had to fill in a questionnaire about their opinion of the system.

In most cases, it took shorter to park using SIAPAS, although the differences in time ranged between 6 and 112 seconds, being the average difference of 28 seconds.

The analysis of the results of the questionnaire showed that all the users thought the system was easy to use, even the ones without experience, but they also agreed in the need to provide a more user-friendly GUI, and specially in the level of detail of the parking representation.

5 Conclusions

Finding a parking space is a common challenge faced by thousands of people every day. Wireless ad-hoc networking technologies offer a new and efficient means to simplify the parking process. In this paper we have described a GPS-based application (SIAPAS) that allows a user to quickly locate and drive to an available parking space.

Our solution is achieved by equipping drivers with a PDA to navigate in the area. The results of the evaluation that has been carried out point out that the system is accurate enough to be useful, which has been confirmed by the people who have taken part in the experiments: in their opinion, the system is useful and easy to use, although some improvements need to be made in the GUI for the application to be a little more user-friendly.

References

1. Mauve, M., Widner, J., , Hartenstein, H.: A survey on position-based routing in mobile ad-hoc networks. *IEEE Network* **15** (2001) 30–39
2. Chon, H.D., Agrawal, D., Abbadi, A.E.: Napa: Nearest available parking lot application. In: *Proceedings of the 18th International Conference on Data Engineering (ICDE'02)*, IEEE (2002) 496–497
3. Basu, P., Little, T.: Networked parking spaces: Architecture and applications. In: *Proceedings of the IEEE Vehicular Transportation Conference*, IEEE (2002) 1153–1157
4. Vicenty, T.: Direct and inverse solutions on the ellipsoid with application of nested equations. *Survey Review* **XXII** (1975) 88–93
5. Cormen, T., Leiserson, C., Rivest, R.: *Introduction to Algorithms*. MIT Press (1990)
6. Juristo, N., Moreno, A.: *Basics of Software Engineering Experimentation*. Kluwer Academic Publishers (2001)
7. Dumas, J.: *A Practical Guide to Usability Testing*. Intellect (1999)

Mobile Social Software for Cultural Heritage Management

Yiwei Cao, Satish Narayana Srirama, Mohamed Amine Chatti, and Ralf Klamma

RWTH Aachen University, Informatik V (Information Systems)
Ahornstr. 55 52056 Aachen Germany
{cao, srirama, chatti, klamma}@cs.rwth-aachen.de

Abstract. In the past several years, the World Wide Web has experienced a new era, in which user communities are greatly involved and digital content explodes via the Internet. Community information systems have been highlighted with the emerging term “Social Software”. In this paper, we explore the impact of social software on the community of cultural heritage management. Furthermore, mobile and ubiquitous technologies have provided capabilities for more sophisticated approach to cultural heritage management. We analyze these features of mobile information systems for cultural communities. We also present a mobile community framework with mobile Web Services to enable professionals to collect, manage and retrieve cultural heritage information in wide user communities.

Keywords: Social Software, Cultural heritage management, mobile Web Services, Web 2.0.

1 Introduction

At the very beginning of the 21st century, a lot of new technologies have emerged. Most of these terminologies always end with *computing*, such as *distributed*, *public*, *grid*, and *social computing*. Among them social computing, also known as *social software*, has been booming due to the simplicity, intuitions and the community base.

Moreover, web technologies have been ready for the second boom after the downfall of the dot-com bubble between 1997 and 2001. Nowadays, a general travel planning scenario should not be alien for any one as an Internet user. Your flight or train tickets are printed at home, since you have booked the tickets online. You get the confirmation message on the cell phone. You surfed at Google earth to have a virtual tour previously and so on.

Impacts of social software technologies have also outreached various communities such as the cultural communities for heritage management. However, the development is still at the beginning stage, initiated by cultural scientists. The quick dissemination in the cultural communities shows the essentials of social software, the sociality. However, the cultural communities have more professional requirements on social software. Thus, we aim at social software to support professionals to research, collaborate and communicate within cultural communities.

The rest of this paper is organized as follows. Section 2 pertains to the impact of social software especially on cultural heritage management. In Section 3, mobile aspects are discussed. A mobile community framework for cultural heritage is proposed. The design and implementation of mobile Web Services which is an innovative technology for the proposed mobile community are introduced in Section 4. Section 5 gives a summary of the paper and discusses the upcoming work.

2 Cultural Heritage Management and Social Software

Cultural heritage is a kind of public goods that includes artefacts and archaeological areas, monuments, group of buildings, single building and the other [24]. We generalize cultural heritage into movable items (artefacts) and geographic heritage (sightseeing). Movable items, also called artefacts, can be preserved and exhibited in museums. So they are generally museum objects.

From the technical point of view, social software technologies are based on Web 2.0. A widely accepted definition of Web 2.0 is an emerging collection of Internet-based social services that provide online collaboration features such as RSS, blogs, wikis, and mashups [22].

2.1 The Impact of Social Software

Lee Bryant has featured social software with smart, simple and social [3]. Web 2.0 - based social software enables communities to collaborate and communicate via the Internet through *smart* idea and *simple* user interface, aiming at *socialization*.

The collaboration of users in online communities ranges from different fields. In the scientific field, the SETI@home Project at the University of California, Berkeley [19] can be seen as the first attempt to involve wide user communities to perform a search for radio signals from extraterrestrial civilizations. In the industrial field, Enterprise 2.0 has emerged [4, 21]. In this, the advocated measures are to move the responsibility of the content manage systems of the companies from administrators to employees' weblogs. Such a bottom-up approach to content organization and delivery is being tested in a controlled experiment at Ernst & Young. 50 employees use Web 2.0 technologies such as blogs and wiki to faster collaboration. Correspondingly, e-learning 2.0 refers to e-learning systems using Web 2.0.

Moreover, in the field of personal information management, personal information and personal activities can find dominating innovative Web 2.0 based social software technologies. Among them are mercora for music, del.icio.us for bookmarks, flickr for images, YouTube for videos, writely for documents, Weblogs for diaries, Google Calendar for calendars, 43things for goals, skype for telephones, instant message for e-mails, meeting friends at MySpace etc. Certainly, there is still some legacy from dot-com era: the Amazon. The most significant feature is *socializing by sharing*. But what is the impact of social software on cultural heritage management?

2.2 The Impact of Social Software in Cultural Heritage Management

The connection between the Web and the community of cultural heritage management is getting tight. The new applications of some social software or Web 2.0 have been

influencing the field of cultural heritage management. For instance, how to create and update an entry in Wikipedia is discussed in [13]. New terminology like Museum 2.0 has emerged [2]. However, the discussion about the applications of Web 2.0 and social computing in museums is solely on how to observe these phenomena and how to use some technologies such as RFID, podcasting and folksonomy [6].

The state-of-the-art research work of Web 2.0 in cultural communities is listed in Table 1. The Steve.museum Project employs social tagging for management of exponents in many museums worldwide. Storytelling has also been employed in several cultural heritage management projects [7]. Since this approach has been used for years, it is hard to evaluate the influence by Web 2.0. Above all, the collaboration feature shows the sociality of Web 2.0. Wide employment of media sharing ideas such as Flickr has not been discovered in cultural communities yet.

Table 1. Web 2.0 technologies in cultural communities

Terms	Web 2.0	Cultural communities
Folksonomy, social tagging	Flickr, delicious	Steve.museum Project (The Metropolitan Museum of Art, Guggenheim Museum, Denver Art Museum, etc.)
Wikis	Wikipedia	Semapedia, Placeopedia
Storytelling	--	Collaborative storytelling
Media sharing	Flickr, Zoomr	--

The quick influence of social software on the cultural communities shows the sociality feature definitely. However, there are still rare cases of social software applications in cultural sites and monuments (sightseeing), such as Google Maps. On the one hand, sightseeing concerns with location information, so that the common information systems can not handle the geographic coordination information well. More complicated geographic information systems are required. On the other hand, mobile technologies may play an important role in providing some location-based services. Consequently, Mobile Social Software (MoSoSo) [9] is highly demanded by user communities. In our work we attempt to explore mobile social software for management of cultural sites and monuments within cultural communities.

3 Cultural Community Goes Mobile with Standards

Usability and sociability are two essential measurements to evaluate online communities [26], which are also key issues for mobile communities. The communities with mobile devices are much larger than the desktop communities.

In this section we discuss the main two approaches. First, the professional cultural community is standardized with metadata. Second, cultural community goes mobile. With both approaches we attempt to define a mobile social software framework for a mobile community of cultural heritage management.

3.1 Standardization with Metadata Standards

Featured with smart and simple, social software is small in its component. However, it should be scalable in user communities of different scales. It should be able to work

with other social software, together accomplishing some complicated tasks. To that end, metadata for description, preservation and administration could play a major role. The wide adoption of RSS feeds by Web sites demonstrates metadata feeds items successfully. At the same time, RSS feeds can be easily syndicated, which proves the concepts of smart and simple of social software.

Two categories of metadata are related to cultural heritage managements: metadata for digital preservation and metadata for cultural heritage. The state-of-the-art metadata standards for digital preservation are systematically surveyed in [8]. The standards are closely associated with some museum-, government- or library-based projects. A comprehensive overview of the related work in this area is reported monthly by the online D-Lib Magazine. Cultural heritage standards include standards for museum objects and location-based sightseeing. Examples of metadata standards are listed in Table 2.

Table 2. Metadata standards in cultural communities

Digital preservation (digital library) [8]	Cultural heritage [17]	
	Museum objects	Cultural sites
ISO OAIS model, MARC, RLG, Dublin Core	CIDOC, Object ID, SPEKTRUM	MIDAS, Core Data Index, Core Data Standard

Many projects and initiatives in cultural communities have developed new standards or extended some existing standards. There are still no dominating metadata standards in cultural heritage management, after decades-long development. Unlike the quick propagation of the Web 2.0 and the social software wave, standardization is a very long tedious process.

3.2 Mobile Communities

Mobile devices have been widely used as digital guides in museums. Moreover, in [23] Headquarter, Mobile Camp and Operative Team build up a hierarchical network. Mobile devices are frequently employed in the level of Operative Team, which collects information on-site. Yet, mobility is still a new topic for cultural communities. The advantages of mobile devices will be increasingly advanced in the aspects of location awareness, one-handed operation, always on and universal alerting device [27].

In addition, recently the capabilities of the wireless devices like smart phones, PDAs are expanding quite fast. This is resulting in their quick adoption in domains like mobile banking, location based services, e-learning, social systems etc.

With these developments, we foster the importance of mobile devices for the cultural communities, based on following reasons. Firstly, more users use cell phones than the Internet over desktops. Secondly, the Internet connection on site might be unavailable sometimes. The professionals can only make use of UMTS, GPRS and the other mobile networks. Thirdly, a distributed system is employed to make backup and replication easily. So the security of the system is enhanced. Finally, usability should not be designed from the viewpoint of the system designers but of communities. User-friendly user interfaces are one of the key points [25].

3.3 Services for Mobile Cultural Heritage Communities

After the discussion about the two approaches above, we aim at designing a mobile community for professionals, using metadata standards. The target groups are professional cultural communities who work on management of cultural sites and monuments. The experiences of developing a desktop-based community information system for cultural heritage management in Afghanistan are also useful [16].

From the technological front, Service Oriented Architecture (SOA) [5] is the latest trend in distributed information systems engineering. Every piece of functionality delivered by any entity in a distributed system can be exposed as a service to the external systems. This concept has been employed in enterprise systems and business processing systems. Web 2.0 and social software are extensions to the SOA concept and can be seen as the first success story of SOA in wide-spread user communities.

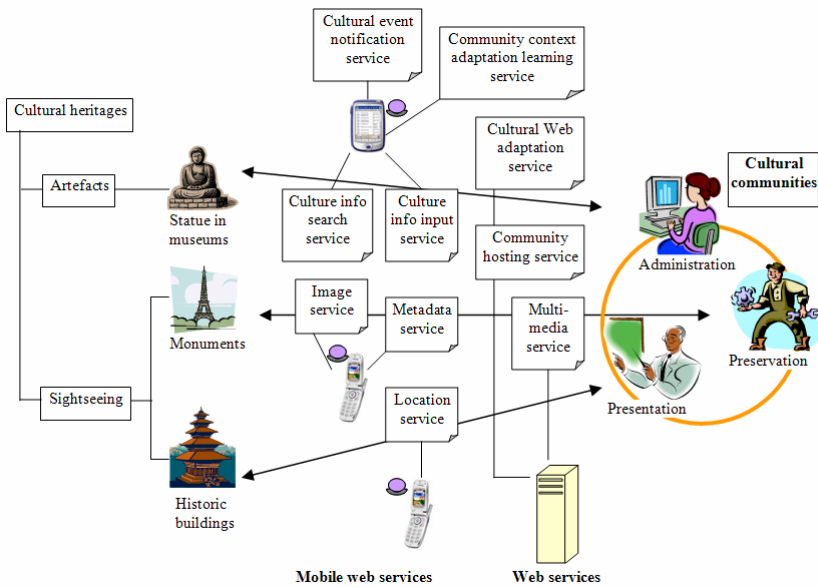


Fig. 1. Professional support for cultural communities through the mobile service framework

The services that are to be provided in a cultural community can be defined as follows (cf. Figure 1). In order to support professionals in cultural heritage management efficiently and flexibly, some services are deployed onto servers, while some services are deployed onto mobile servers such as cell phones and PDAs.

In case that the services need more computing and storage capacities, they can be realized with usual Web Service technologies [32] on work stations or servers. For example, a *cultural Web site adaptation service* needs to be defined to adapt usual cultural Web site for cell phones and PDAs according to the display capacities. *Multimedia services* are to be responsible for multimedia management, e.g.

multimedia search on site of historic buildings and other sightseeing. A *community hosting service* is used for user management including new users registration, log in and user profile management, and multimedia access right management (cf. [28]).

The rest services based on cultural heritage or communities can be deployed on mobile servers. In the aspect of cultural heritages, there are mainly *cultural information input services*, *cultural information search services*, and *metadata services*. A *cultural information input service* provides professional users to input data with selected standards according to users' wish. A *cultural information search service* enable users to search site descriptions in text. Since there are lots of standards in the field of cultural heritage management and most of the standards are based on texts, the file size of the application is not very big. A *metadata service* can be defined to do mappings among different standards.

In the aspect of cultural communities, the defined services are *location service*, *cultural event notification service*, and *community context adaptation learning service*. A *location service* can be used to locate professionals at fieldwork. A *cultural event notification service* is to send alerts onto mobile devices to inform professionals about cultural events as well as the presence of the community members in the neighbourhoods. A *community context adaptation learning service* is deployed to deliver professionals learning stuff with regard to the community context. This use case is proposed in detail in [18]. An *Image service* is provided for picture upload.

These services can be mainly accessed by the mobile devices. To realize these requirements, we employ an approach of mobile Web Services [29], which will be introduced systematically in the next section. With mobile Web Services the mobile devices can participate in service consumption as well as service delivery.

3.4 Summary of Mobile Cultural Communities

Summarily, the benefits to employ mobile Web Services are, on the one hand, the flexibility and that service provider and service consumer can be on the same devices. On the other hand, the deployment can be executed via mobile devices, if the mobile network is available. However, there are several potential problems and challenges. First, the instability exists. Services might be easily removed by the mobile device owner. Next, the capacity is still quite limited despite the rapid development of hardware. Last, it lacks a business model to control the charge of the services.

Technical companies together with W3C have agreed on mobile Web rules lately, in order to solve the problems of low visits of mobile Web sites. Google has also just launched its mobile personalized site in Europe. Although all these measures provide some soft conditions, it is still hard to let user communities pay for mobile content services as willingly as for phone calls.

However, our mobile community framework tries to employ new technologies to promote more mobile use cases. To support such mobile social software, mobile Web Services alone can not meet all requirements. The framework using usual as well as mobile Web Services can make good use of the advantages of both services. Thus, such a piece of mobile social software can perform the tasks more efficiently and flexibly.

The following section explains the details of mobile Web Services and the realization details of social software services. The discussion has to get into some technical detail, but we believe that this kind of approach can also carry over to many other of the forthcoming pervasive applications of mobile information systems in social networks and may therefore be worthwhile presenting here.

4 Mobile Web Services

Service Oriented Architecture is the latest trend in information systems engineering. It is a component model, presenting an approach to building distributed systems. SOA delivers application functionality as services to end-user applications and other services, bringing the benefits of loose coupling and encapsulation to the enterprise application integration. A service having a neutral interface definition that is not strongly tied to a particular implementation is said to be loosely coupled with other services. SOA is not a new notion and many technologies like CORBA and DCOM at least partly represent this idea. Web Services are newest of these developments and by far the best means of achieving SOA.

The Web Service architecture defined by the W3C enables application-to-application communication over the Internet. Web Services are self-contained, modular applications whose public interfaces are described using Web Services Description Language (WSDL) [33]. Web Services allow access to software components through standard Web technologies and protocols like SOAP [34] and HTTP [12], regardless of their platforms, implementation details. A service provider develops and deploys the service and publishes its description and binding/access details (WSDL) with the UDDI registry [31]. Any potential client queries the UDDI, gets the service description and accesses the service using SOAP. [10] The communication between client and UDDI registry is also based on SOAP.

Web Services and its protocol stack are based on open standards and are widely accepted over the internet community. Web Services have wide range of applications and range from simple stock quotes to pervasive applications using context awareness like weather forecasts, map services etc. The biggest advantage of Web Services lies in its simplicity in expression, communication and servicing. The componentized architecture of Web Services also makes them reusable, thereby reducing the development time and costs.

The quest for enabling these open XML Web Service interfaces and standardized protocols also on the radio link lead to new domain of applications mobile Web Services. In this domain, the resource constrained mobile devices are used as both Web Service clients and providers. Figure 2 shows the deployment scenario of mobile Web Services, where mobile devices are used as both Web Service providers and clients. While mobile Web Service clients are quite common these days, the research with mobile Web Service provisioning is still sparse. To support this, during one of our previous projects, we have developed and analyzed the performance of a mobile Web Service provider on smart phones. [1, 15, 29]

Mobile Host is a lightweight Web Service provider built for resource constrained devices like cellular phones. It has been developed as a Web Service handler built on top of a normal Web server. The Web Service requests sent by HTTP tunneling are diverted and handled by the Web Service handler. Using HTTP tunneling it is

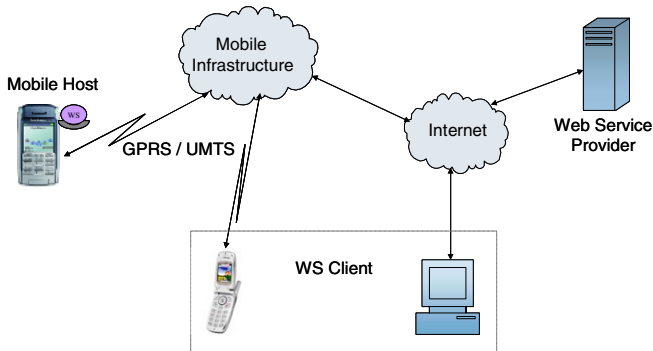


Fig. 2. Mobile Web Services scenario

possible to send data of any protocol through proxy over HTTP. The protocol messages are wrapped into the HTTP message body and are transferred as normal HTTP GET/POST requests. Detailed description of Mobile Hosts' design is beyond the scope of this paper and is available at [29].

The Mobile Host was developed in PersonalJava [14] on a SonyEricsson P800 smart phone. The footprint of our fully functional prototype is only 130 KB. Open source kSOAP2 [20] was used for creating and handling the SOAP messages.

The detailed evaluation of this Mobile Host clearly showed that service delivery as well as service administration can be done with reasonable ergonomic quality by normal mobile phone users. As the most important result, it turns out that the total WS processing time at the Mobile Host is only a small fraction of the total request-response time (<10%) and rest all transmission delay. This makes the performance of the Mobile Host directly proportional to achievable higher data transmission rates. Thus, the high data transmission rates achieved, in the order of few Mbps, through advanced mobile communication technologies in 2.5G, 3G and 4G, help in realizing these Mobile Hosts in the commercial applications [11, 35]. The Mobile Host was also successful in handling concurrent accesses for reasonable service like location data provisioning service.

Mobile Host opens up a new set of applications and it finds its usage in many domains like collaborative learning, social systems, mobile community support and etc. Many applications were developed and demonstrated using Mobile Host, for example in a distress call, the mobile terminal could provide a geographical description of its location along with location details. Similarly interesting scenarios like mobile expertise finder services, mobile learning media sharing services are possible in e-learning domain. Current research in this domain concentrates on adopting this Mobile Host feature into the social software systems. [30]

5 Conclusion and Outlook

The mobile community support for cultural heritage researchers is still in its early phase. The impact of social software on professional communities in various fields is

significant. It is time to survey the adaptation of the requirements and design of social software esp. mobile social software from non-professional level to professional level.

In this paper we made an analysis of the relationships between social software and cultural communities. Standardization with metadata and mobility are main issues to provide professionals a mobile community for cultural heritage management. We have defined the basic services to support the community. The basics of mobile Web Services technology are also introduced. The upcoming tasks are to apply the mobile Web Services into the designed mobile community framework for cultural heritage management.

References

- [1] N. Balani: Deliver Web Services to mobile apps. IBM developerWorks (2003)
- [2] D. Bearman: Museum 2.0, Museums and Web 2006 the international conference for culture and heritage on-line, Albuquerque, New Mexico (2006)
- [3] L. Bryant: Smarter, Simpler Social - An introduction to online social software methodology, v. 1, Headshift, April 2003, <http://www.headshift.com/moments/archive/sss2.html>,
- [4] L. Bryant: Humanizing the Enterprise using Ambient Social Knowledge, O'Reilly Emerging Technologies Conference, March 5-9, 2006, San Diego, CA, USA.
- [5] S. Burbeck: The tao of e-business services: The evolution of web applications into service-oriented components with Web Services, IBM Developerworks, <http://www-106.ibm.com/developerworks/webservices/library/ws-tao/> (2000)
- [6] S. Chun, R. Cherry, D. Hiwiller, J. Trant, B. Wyman: Steve.museum: An Ongoing Experiment in Social Tagging, Folksonomy, and Museums, Museums and Web 2006 the international conference for culture and heritage on-line, Albuquerque, New Mexico (2006).
- [7] T. Collins, P. Mulholland, D. Bradbury, and Z. Zdrahal: Methodology and Tools to Support Storytelling in Cultural Heritage Forums, *Proceedings of the 14th international Workshop on Database and Expert Systems Applications*, September 01 - 05, 2003, DEXA. IEEE Computer Society, Washington, DC.
- [8] M. Day: Metadata for Digital Preservation: A Review of Recent Developments, P. Constantopoulos and I.T. Sølvyberg (eds.): *Research and Advanced Technology for Digital Libraries*, Proceedings of 5th European Conference, ECDL 2001, Darmstadt, Germany, September 2001, Springer LNCS 2163.
- [9] N. Eagle, A. Pentland: Social serendipity: mobilizing social software, *Pervasive Computing* Vol. 4, Nr. 2, IEEE (2005)
- [10] K. Gottschalk, S. Graham: Introduction to Web Services Architecture, *IBM Systems Journal* 41(2): 178-198, (2002)
- [11] GSM World: GSM - The Wireless Evolution, (2006) <http://www.gsmworld.com/technology/index.shtml>
- [12] HTTP: Hypertext Transfer Protocol version 1.1. IETF RFC 2616, <http://www.ietf.org/rfc/rfc2616.txt> (1999)
- [13] Jonathan Bowen, Jim Angus: Museums and Wikipedia, Museums and Web 2006 the international conference for culture and heritage on-line, Albuquerque, New Mexico (2006)
- [14] JCP: PersonalJava application environment specification, version 1.2a. SUN Developer Network, <http://java.sun.com/products/personaljava/index.jsp>, (2000)
- [15] JSR 172: J2ME Web Services Specification. Java community process.

- [16] R. Klamma, M. Spaniol, M. Jarke, Y. Cao, M. Jansen, G. Toubekis: A Hypermedia Afghan Sites and Monuments Database, *Proceedings of the First International Workshop on Geographic Hypermedia "Geographic Hypermedia: Concepts & Systems"*, Denver, USA, April 4-5, 2005, pp. 59-73.
- [17] R. Klamma, M. Spaniol, M. Jarke, Y. Cao, M. Jansen, G. Toubekis: Standards for Geographic Hypermedia: MPEG, OGC and co., *N.T. Jepsen et al. (eds): Geographic Hypermedia: Concepts & Systems*, Springer (2005)
- [18] R. Klamma, M. Spaniol, Y. Cao: Community Aware Content Adaptation for Mobile Technology Enhanced Learning, (to be appear in) *Proceedings of the First European Conference on Technology Enhanced Learning*, Crete, Greece, October 1-4, 2006.
- [19] E. Korpela, D. Werthimer, D. Anderson, J. Cobb, and M. Leboisky: SETI@home-massively distributed computing for SETI, *Computing in Science & Engineering*, Vol. 3, No. 1, pp. 78-83, IEEE (2001)
- [20] kSOAP2: A open source SOAP implementation for kVM, <http://kobjects.org/> (2006)
- [21] A. McAfee: The Trends Underlying Enterprise 2.0, Harvard Business School Faculty Blog, (2006) available at http://blog.hbs.edu/faculty/amcafee/index.php/faculty_amcafee_v3/the_three_trends_underlying_enterprise_20/
- [22] J. McKendrick: Web 2.0 or SOA? Web 2.0 and SOA? Let the Debate Begin! – Part 1, [webservices.org](http://www.webservices.org/), (2006) available at http://www.webservices.org/weblog/joe_mckendrick/web_2_0_or_soa_web_2_0_and_soa_let_the_debate_begin_part_1
- [23] A. Maurino, S. Modafferi: Challenges in designing of cooperative mobile information systems for the risk map of Italian cultural heritage, *Proceedings of the 4th International Conference on Web Information Systems Engineering Workshops*, IEEE (2004)
- [24] S. Navrud, R.C. Ready (eds.): *Valuing Cultural Heritage – Applying Environmental Valuation Techniques to Historic Buildings, Monuments and Artifacts*, Edward Elgar Publishing Ltd., UK (2002)
- [25] J. Nielsen: *Usability Engineering*, Clarendon Press (1993)
- [26] J. Preece: *Online Communities: Designing Usability, Supporting Sociability*, John Wiley and Sons (2000)
- [27] Jo Rabin, Charles McCarthieNevile (eds.): *Mobile Web Best Practices 1.0 – Basic Guidelines*, W3C Working Draft 18 May 2006, <http://www.w3.org/TR/mobile-bp/>.
- [28] M. Spaniol, R. Klamma, M. Jarke: ATLAS: A web-based software architecture for multimedia e-learning environments in virtual communities, *W. Zhou, P. Nicholson, B. Corbitt, J. Fong (Eds.): Advances in Web-Based Learning, Proceedings of ICWL 2003*, Melbourne, Australia, August 18-20, 2003, Springer-Verlag, Berlin Heidelberg, LNCS 2783, pp. 193-205.
- [29] S.N. Srirama, M. Jarke, W. Prinz: Mobile Web Service Provisioning, *Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/Advanced*, Feb. 19-25, 2006, pp.120 – 126.
- [30] S.N. Srirama, M. Jarke, W. Prinz: Mobile Host: A feasibility analysis of mobile Web Service provisioning. In: *4th International Workshop on Ubiquitous Mobile Information and Collaboration Systems (UMICS 2006)*, a CAiSE'06 workshop (2006)
- [31] UDDI: The Universal Description, Discovery and Integration, <http://www.uddi.org/> (2004)
- [32] W3C: Web Services Activity, <http://www.w3.org/2002/ws/>, May 2004.
- [33] W3C: WSDL, Web Services Description Language, version 1.1, (2004) <http://www.w3.org/TR/wsdl>
- [34] W3C: SOAP, Simple Object Access Protocol, v. 1.1, <http://www.w3.org/TR/SOAP> (2004)
- [35] 4G Press: World's First 2.5Gbps Packet Transmission in 4G Field Experiment, <http://www.4g.co.uk/PR2006/2056.htm>, (2005)

Middleware Platform for Ubiquitous Location Based Service

Jae-Chul Kim, Jai-Ho Lee, Ju-Wan Kim, and Jong-Hyun Park

Telematics & USN Research Division,
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
{kimjc, snoopy, juwan, jkp}@etri.re.kr

Abstract. Over the past few years, several studies have been made on Location Based Service (LBS) middleware platform about performance and architecture. And service area is extended as mobile client is an indispensable factor in LBS. Although LBS middleware platform process a large transactions to cope with interoperability and real time processing, current LBS system use an existing Geographical Information System (GIS). In this paper, we propose an Open LBS Middleware Platform (OLMP) which is possible to process a large moving object and to support different mobile clients such as PDA and cellular phone. We describe the system architecture of an OLMP and a main memory DBMS. A proposed OLMP is consisted of open LBS components, mobile gateway, and main memory DBMS.

1 Introduction

Location Based Services encompass broad areas of technology and the need for real-time interoperable processing is evident for engagement between technology components. Such components include Content & Applications, Gateways & Middleware, Network Equipment, Service Provider, and End-user Devices. Seamless integration of these into a reliable Location based service requires open interface for data exchange and real-time data processing. But Conventional methods of proprietary data and components have distinct disadvantages in terms of interoperability and integration and do not allow much flexibility in cost and scalability [9].

In the age of information explosion and technological advancement, ubiquitous location based service is becoming a significant feature in the era of telecommunication [1]. The ubiquitous location based service is a requirement for certain telecommunication or mobile applications that uses location information. This development is linked to the tremendous growth in the number and the sophistication of mobile phone and mobile technology. And the trend continues stealthily invading mobile domains especially of those that utilize geographical positions or location information of the mobile devices or that of the mobile user. Various ubiquitous location based service applications that are available in the market are normally tailored to a specific technology. Most of these applications require support from a combination of a number of

technologies such as location sensor technologies (GPS, MSR RADAR, etc) and service providers [2].

The examples of LBS are a buddy finder service to find a location of a friend, a navigation service to provide routing information to a driver, L-commerce to advertise goods based on customer's location, and "E911" service for emergency calls. In those applications, there are enormous numbers of "moving objects" to be managed and queried. Since the moving objects may report their locations frequently, a database system should be able to handle a huge number of updates and queries quickly [3]. However, traditional disk resident relation DBMS cannot handle the updates and queries efficiently. Even more it does not support a query language specific to handling moving objects. It means application developer should concern all the things to update and retrieve the locations.

To solve those problems simultaneously, architecture of middleware should assure interoperability and assist various mobile clients. Moreover, it is needed that a large moving object and GIS data should be processed in real-time.

In this paper, we proposed an open LBS middleware platform architecture that guarantees interoperability and real-time processing. It was experimented in within the South Korea. Section 2 describes a general overview of the LBS platform architecture. The proposed prototype implementation is described in Section 3; followed by a conclusion and future works.

2 Proposed Open LBS Middleware Platform Architecture

2.1 System Architecture

As Location based service provider is not dependent on the wired (or wireless) network companies, some commercial LBS platform should have characteristics of interoperability, flexibility and scalability. Those products are MapPoint Location Server(MLS), QUALCOMM Internet Services (QIS), Hewlett Packard's OpenCall MLS, and Openwave Location Services Platform [5]. But these platforms do not consider the standard specification and real-time processing of the large moving objects.

Moreover, LBS Solution Company and Developer have a burden on system integration. To solve these problems, the proposed platform is composed of XML web service using open (standard) interface and main memory DBMS to process moving object and GIS data. And mobile gateway is included to support a client which can not process XML data.

The proposed Open LBS Middleware Platform (OLMP) is described in figure 1. Each sub-system is composed of an open LBS components (travel advisory, routing, presentation, location utility, directory, tracking, positioning component), mobile gateway, and main memory DBMS. And positioning service is connected with MPC (Mobile positioning Center) of telecommunication companies (SKT, KTF, LGT : Korean mobile companies). Information of travel advisory service is provided by traffic, road construction information, and weather Information Company.

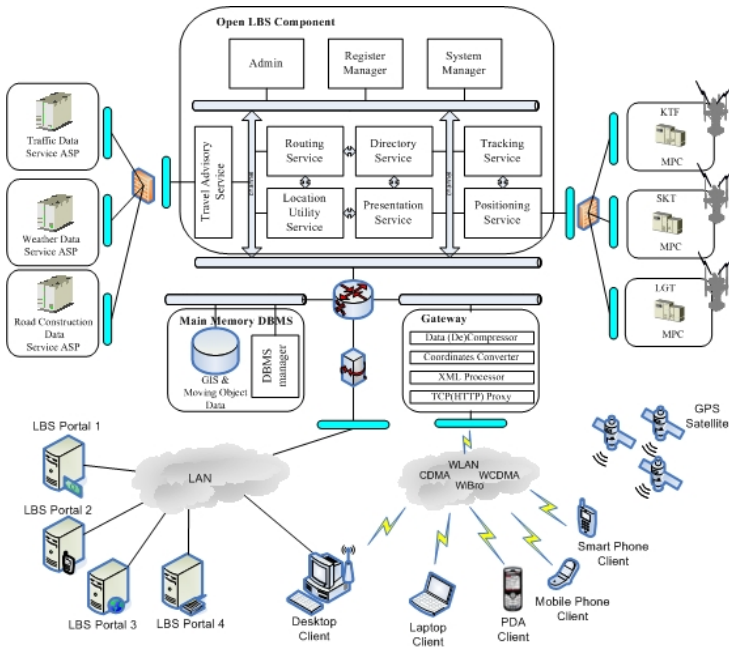


Fig. 1. System architecture

The interface is developed according to Open Mobile Alliance(OMA), Open Geospatial Consortium (OGC), and Telecommunication Technology Association (TTA).

2.2 Open LBS Component

Recently, middleware platform adopts web service architecture to communicate with other middleware platform. One of the newest innovations for the use of the Internet is web services. Web services allow applications and Internet-enabled devices to easily communicate with one another and combine their functionality to provide services to each other, independent of platform or language. Web services are characterized by SOAP messages used to talk to a web service, WSDL files that describe a web service, and the UDDI used to find web services. Conceptually, web services are very understandable. They eliminate many of the complexities that have been required when there is a need for computer applications to interact with each other.

Open LBS components (routing , presentation , directory, location utility, travel advisory, tracking, positioning service) is working in multi-application server and connected with one or a few DBMS. Here, application server is working with rule of round robin.

2.3 Mobile Gateway

Although XML web service is interoperable, XML data, too huge burden to cellular phone, is not adequate to mobile client. Because most mobile client platform has not a

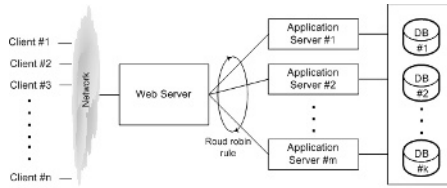


Fig. 2. System model of open LBS components

XML parser module, XML response data from LBS server based on web service can not be interpreted in mobile device. These problems force the LBS provider company to have dual LBS server for interoperable service. But, in the proposed architecture, these drawbacks are solved by using Mobile Gateway. In figure 3, mobile gateway is composed of server proxy module, request/response adaptor, data (de)compressor, and coordinate converter.

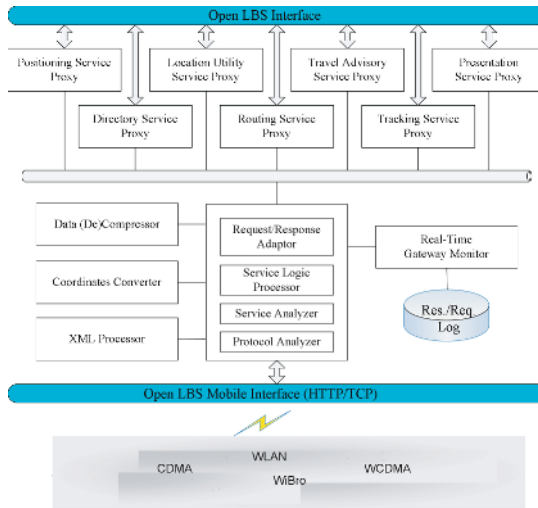


Fig. 3. Structure of gateway

Protocol Analyzer interpret the communication protocols (HTTP and TCP). And Service Analyzer classifies the request from mobile client according to each service type. Subsequently, Service Logic Processor interprets request type, and Request Adaptor transform request type into XML request data. Server proxy modules (Presentation Service Proxy, Directory Service Proxy, Location Utility Service Proxy, Routing Service Proxy, Travel Advisory Service Proxy, Tracking Service Proxy, and Positioning Service Proxy) are interface proxies that are connected to OLMP (Open LBS Middleware Platform) for exchanging the request and response message.

2.4 Main Memory DBMS

When tracing wireless network users, tracking fleet vehicles, finding the best way to deliver goods and services, or analyzing transportation traffic, resolving the problems from a moving objects perspective is crucial to providing advanced location-based services [10]. Traditional database systems have two major problems in managing moving objects. One is that conventional disk-based database systems cause disutility to cope with massive update operations of location information. The other is that, since they don't support moving objects data model and query language, application developers should implement all of them.

Figure 4 shows the overview architecture of the main memory moving objects database system and applications. Moving objects databases consists of moving objects components, moving objects SQL processor, and a main memory storage including moving objects indexes.

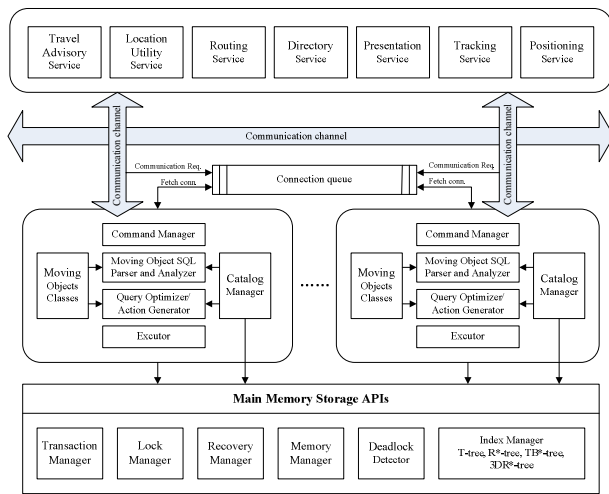


Fig. 4. Architecture of Moving Object Database System

Moving objects class components are a set of temporal, geometry, and moving objects classes that implement moving objects data model and operations. Moving objects SQL processor would processes and executes various and powerful moving objects queries. The extended query language is compatible with SQL3, and supports to make application system like moving objects data mining and customer relationship management through application programming standard such as JDBC, ODBC, and OLEDB. Main memory storage component consists of transaction manger, lock manager, recovery manager, deadlock detector, and index manager as like other conventional DBMS. But, we support current memory location indexes and past moving objects index such as R*-tree [6], TB*-tree [7], and 3DR-tree [8]. This will increase the performance of moving object access efficiently.

2.4.1 Modeling Moving Objects

Temporal classes consists of Period, Interval, Instant, and TemporalCollection classes(see figure 5). These classes have interfaces ITemporal, ITemporalRelation, and ITemporalOperator Interfaces.

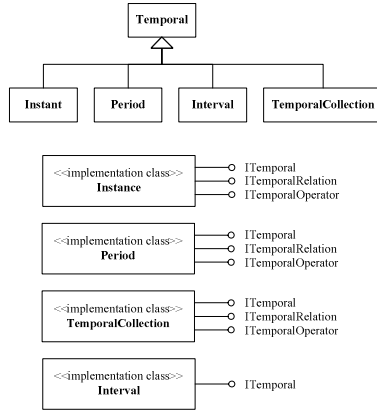


Fig. 5. Temporal Class Hierarchy

Geometry classes consists of Point, LineString, Polygon, GeometryCollection, Surface classes, and et al. UML modeling for geometry classes is borrowed from its of International Standard of Open GIS Consortium [4] for geographic information system. Classes for Moving objects consists of TObje, MObject, MBase, MGeometry classes, and et al (see figure 5 and figure 6).

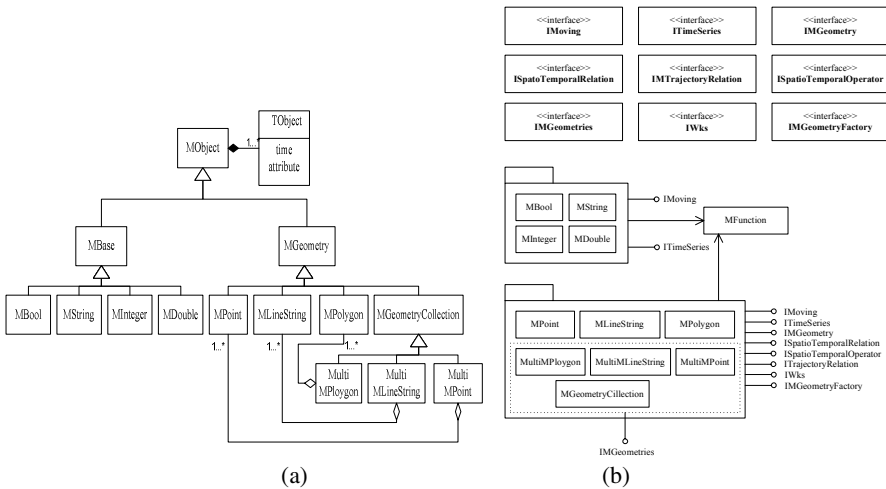


Fig. 6. (a) Moving Objects Class Hierarchy (b) Interfaces and Classes for Moving Objects

3 Implementation

A proposed system is tested using PC, PDA, and mobile phone client (see figure 7, figure 8, and figure 9). Mobile client, connected to mobile gateway, is tested in each service case, but PC and PDA client, connected to open LBS component or mobile gateway, is tested in each service case twice. Each system specification is described in table 1.

Table 1. Client specification

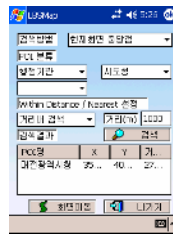
	OS	CPU	Memory	Network	GPS	Model
PC	Windows XP	Intel Pentium 4 3.0Ghz	512MB	LAN	None	LG XPION 260
PDA	Windows Mobile 2003 SE	Intel PXA272 Bulverde	128MB (ROM) & 64MB (RAM)	CDMA2000 1x EVDO & 802.11b	External GPS	Samsung SPH-M4300
Mobile Phone	Wireless Internet Platform for Interoperability (WIPI)	MSM600 0 series	128MB	CDMA2000 1x EVDO	Standalone GPS & A-GPS	Samsung SPH-S1100



(a) Travel Advisory service



(b) Route service



(c) Directory service



(d) Tracking service



(e) Location Utility service



(f) Presentation service



(g) Positioning service

Fig. 7. PDA client

Overall test-result tell us that the proposed system could be used in (wireless) LAN or Code Division Multiple Access (CDMA) network with PC, PDA, and mobile phone.

Figure 4 illustrates the screens displayed on the PDA(Samsung SPH-M4300). Each screen shot implies the result of service request. All processes in this PDA system are carried out in server side and just a response in XML form (or binary data form Mobile Gateway) is sent to the PDA Client. In case of PDA client, XML response from LBS Middleware Platform and binary data from Mobile Gateway is possible to be processed.

Figure 5 illustrates the screens displayed on the mobile phone (Samsung SPH-S1100). In this case, each request and response is processed by way of Mobile Gateway. XML parser in Mobile Gateway interprets the XML contents and a mobile phone receives a binary data from Mobile Gateway to display the result.

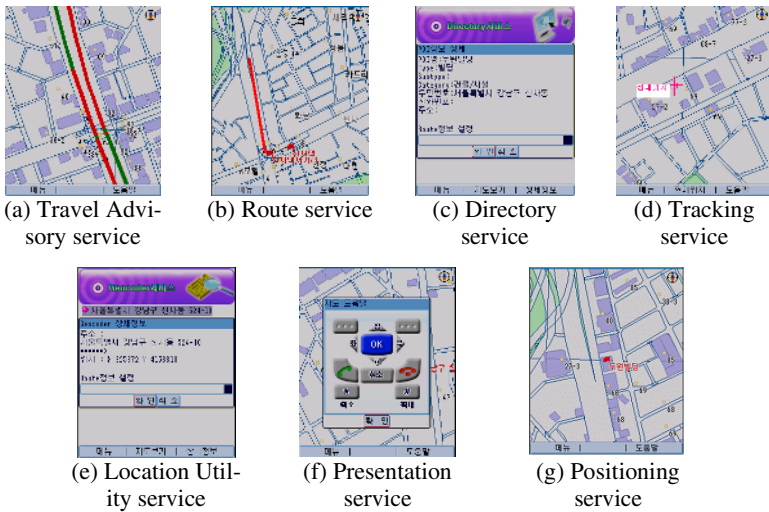


Fig. 8. Mobile phone client

4 Conclusion and Future Work

Current LBS platform has a problem on interoperability and real-time process of large moving objects (GIS data). In this paper, we proposed an Open LBS Middleware Platform (OLMP) using main memory DBMS to overcome the limitation of interoperability and data process time. The proposed OLMP provides an strong contrast to other commercial LBS middleware platform with mobile client independency and real-time processing of 3D(2D geometry(object) and time) data.

The future study focuses on a personalized LBS system for context awareness.

References

1. Hightower J. and Borriello G. : Location Systems for Ubiquitous Computing. IEEE Computer Society, Vol. 34 Num. 8, (2001) 57-66
2. Seong-Baek Kim, Kyung-Ho Choi, Seung-Yong Lee, Ji-Hoon Choi, Tae-Hyun Hwang, Byung-Tae Jang, and Jong-Hun Lee, "A Bimodal Approach for Land Vehicle Localization," ETRI Journal, vol. 26, no. 5, Oct. 2004, pp.497-500
3. Coteló L., J. A. Forlizzi,, L. Guting, R. H. , Nardelli, E., and Schneider, M.: Algorithms for Moving Objects Databases. FernUniversität Hagen, Informatik-Report 289, October (2001)
4. Open GIS Consortium, Inc.: OpenGIS Simple Features Specification For OLE/COM Revision 1.1 (1999)
5. U.S. Location-based Service (LBS) Markets-Defining the Enterprise Opportunity F134-65. Frost & Sullivan (2005)
6. Brinkhoff , Hans-Peter Kriegel , Bernhard Seeger : Efficient processing of spatial joins using R-trees. Proceedings of the 1993 ACM SIGMOD international conference on Management of data. (1993)
7. Dieter Pfoser, Christian S. Jensen, Yannis Theodoridis: Novel Approaches in Query Processing for Moving Object Trajectories. VLDB 2000 .
8. Vazirgiannis, M., Theodoridis, Y., and Sellis, T.: Spatio-Temporal Composition and Indexing for Large Multimedia Applications. Multimedia Systems. (1998)
9. LBS The Ingredients and The Alternatives, Last Access Date 24th October 2005, at url: <http://www.gisdevelopment.net/technology/lbs>.
10. Pankaj K. Agarwal, Lars Arge, and Jeff Erickson. : Indexing moving points. In Proceedings of the ACM Symposium on Principles of Database Systems (PODS). (2000)

Multimodal Architectures: Issues and Experiences

Giovanni Frattini, Luigi Romano, Vladimiro Scotto di Carlo, Pierpaolo Petriccione,
Gianluca Supino, Giuseppe Leone, and Ciro Autiero

AtosOrigin Italia S.p.A., Via Antiniana 2/A, 80078 Pozzuoli (NA), Italy
{giovanni.frattini, luigi.romano, vladimiro.scottodicarlo,
pierpaolo.petriccione, gianluca.supino, giuseppe.leone,
ciro.autiero}@atosorigin.com

Abstract. The penetration of mobile device in western countries is still increasing. The Italian case is really surprising: every single Italian has more than one mobile terminal. Thus, considering this large potential audience, there is real need for innovation and new services. In this context, usable multimodal services could have an unexpected impact on social behaviour. Nevertheless, the research community should be able to propose a framework for building generic multi-modal services, covering all their lifecycle. We are currently defining an architecture for building coordinated simultaneous multimodal applications trying to use as much as possible open source software: our goal is to define a set of tools for enabling a rapid deployment of a generic multimodal service. In our opinion, a platform based on open source software could meet the expectations of a large numbers of service developers. A special effort for enabling a mass diffusion of mobile multimodal services should be focused on the client side, where the situation is still evolving.

1 Introduction

Atos Origin Italy and the Italian Ministry of Research have funded a new project called C.H.A.T. which main purpose is to study multimodal systems using mobile terminals. When we have started this project we found that, in recent years, many papers have talked about the great advantages we should have had interacting with a multimodal computer interface [1,2,3,4,8]. Some authors have studied the usability of a multi-modal interface for common users, trying to establish how multimodality could add a real value for them [10,17]. Oviatt's *Ten Myths of Multimodal Interacion* [10] remains a pillar of any research on multimodality since it shows clearly the real nature of multimodal interaction.

In the meanwhile, on the market we are starting to see the first multimodal applications (Kirusa is one of the most active companies, see <http://www.kirusa.com>). Such applications are though for being used on mobile terminals. However, even if some attempt to open the market of mobile multimodal applications is in progress, we believe that the target should be reconsidered: most of the application both commercial or experimental, are based on very powerful PDA and on .Net technologies (Microsoft). This is a good choice for addressing high-end terminal and a small slice of the potential users (and, in fact, some interesting application exists). In

the meanwhile the penetration of Java enabled phones makes even more interesting to reconsider multimodality for this huge market. On the other hand, some of the bigger players on the market (Google, Opera) are distributing Java applications for mobile handsets, and, thus, there are important sign that this is the market direction . Thus, the possibility of thinking to a multimodal platform that could be used for enabling not just PDAs, but also smart-phone and java-phone become even more interesting.

The above discussion on terminals and multimodality could be generalised in some extent. If the “scalability” model on the client is still an issue, the same could apply for the server. We have considered a lot of papers reporting very interesting and scientifically relevant results, but few of them are trying to move in the direction of an effective engineering of a multimodal platform and, from our point of view, this is one of the major barriers for a realistic diffusion of multimodal application on the market: creating a new multimodal application (at least a synergic multimodal application) is still a research lab matter, where very high-skilled people are able to collect, aggregate and develop all the necessary software. There are many issues that must be still faced and solved and, thus, there is still a lot of work to do and several open issues and that is why we want to open a discussion on them.

2 Open Issues for a Practical Multimodality

Let us start from considering what are for us the priorities:

- Define a development methodologies for ensuring the highest level of usability. It is clear from [3] that multimodality is not a panacea for user interaction. Usability must be considered very well. This is even more true in the context of small devices. Thus a methodology that concentrate the attention on usability [18] and ensure it all along the development cycle is fundamental
- Define a evolution model for multimodal platforms. Most of the potential multimodal service providers are currently just “service providers”. In other words they are now offering services using “traditional” multi-channel service delivery platforms. It is quite obvious that the evolution toward multimodal services should save as much as possible their investment. It is not neither practical nor effective to move toward technologies that are proprietary or even not “well-known”. The idea we are pursuing is to build modules for enabling multimodality on top of standard middleware, trying to ensure the minimum impact on pre-existent service deliver platforms.
- Consider as much as possible very small terminal and their limited capability for enabling multimodal services (even if very simple) starting from the bottom going up. The support of MExE [15] platform is a fundamental requirement for penetrating the market and, thus, contribute to a real change of the people lifestyle.
- Thin/Fat client architecture. A framework for building multimodal services should be able to support both fat and thin clients. A constraint on clients could lead to a new barrier to the multimodality diffusion. When speaking about thin vs fat client we do not refer simply to a browser mediated interaction but more in generally to the physical location of the multimodal modules. PDAs and smart-phone are able to run speech recognition applications (see, for example, www.nuance.com) and

TTS application, while other mobile devices are not enabled to run anything but Java application. Thus, a platform for building a generic multimodal application should be able to receive raw data and pre-processed data, depending on the terminal capabilities.

- Independence from the recognisers. Considering that the W3C activities and standard are concentrated on how to represented data coming from the recognisers ([11]), it is not so obvious how to transport and inject data into a generic recogniser. An integration module for ensuring the independence from the specific technology adopted should be in place.
- Being open to new standard. It is a matter of fact that does not exist a standard for building synergic multimodal application at least on the client side. While X+V[14], EMMA [12], SGML [13] and all the activities of the W3C around multimodality are addressed to create standards for multimodal interactions, for synergic multimodality we have not found any useful initiative. Being open to new standard means to work around software modules that could converge easily toward a new standard.
- Consider as much as possible the user context. We think that a multimodality and context-aware computing should be considered as two faces of the same medal: acquire information of the user context could enrich the information for a better understanding of the user willing and work together to a better and effective user experience.
- It's interesting to note that, at beginning, the efforts on multimodal integration have concentrated mainly on semantic representations and incorporation of new input technologies [6,8,9] while, later, other studies started considering the statistical integration process that defines a multimodal system architecture [19]. In any case, a modular extensible framework could add several benefits. For example it could be particularly useful to add or remove modality analyzers or renderer components to adjust the system's capabilities to different user contexts, implementing a plug and play architecture [20].

It is important to underline the aspects related to a lack of standards. The few companies that are on the market with real multimodal application on mobile clients, are using proprietary technologies (and they are not giving any detail on them). Languages like X+V and SALT are too much form-centered and are not powerful enough for a real synergic multimodality. These facts impose strong constraints: whatever you are going to develop now must be based on a independent initiative. The problem thus could be stated: how can I build a framework that could be easily expanded or refactored for supporting new eventual standards (especially on mobile terminals). This framework should necessarily be modular and support the following features: sensitive screen (if any), simple management of input channels, integration of input and output in the same interface.

The third point is especially crucial and impact the way content are presented to the user. Again, does not exist a standard for presenting multimodal content on client terminals even if an extension of SMIL [15] (SMIL/ReX) could be a good candidate [7]. Anyway it should be not simple to integrate in the same interface a SMIL output with input modules in a single coherent interface.

3 Our Reference Architecture

The proposed architecture, showed schematically in Fig. 1, is composed of different modules. As discussed in the introduction we have analysed different papers concerning multimodal architectures and platforms. Our reference architecture follows the general guidelines described in these papers, especially [4,24,25,26].

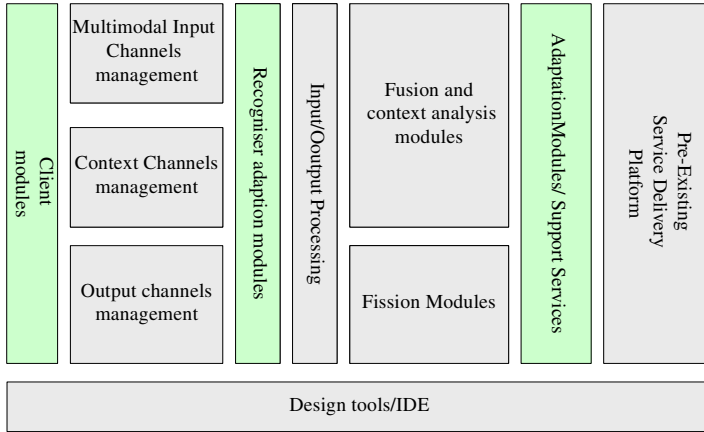


Fig. 1. Our Reference Architecture

Nevertheless, we would put the attention on aspects that are not discussed in depth in literature (highlighted in the figure):

- **Recogniser independence.** There are several product on the market that could be used for executing recognition (speech recognition, handwriting recognition). For integrating them into a multimodal architecture their capabilities should be standardised. Nevertheless, we have not found any trace of such an effort. This mean that a system using a specific recogniser could be tied to it forever. For such a reason we believe that a discussion on standardising recogniser interfaces could help in building new generation multimodal services. In the meanwhile, a software layer for abstracting recognisers from their real implementation is more than necessary: this software layer could enable the reuse of existing recognisers (especially for speech recognisers).
- **Adaptation modules.** Most of the interest of the community is currently concentrated in building new multimodal platforms and interfaces. We have not found, however, any discussion on how to reuse Service Delivery Platforms that are up and running in hundreds of companies around the world. The best strategy for achieving the complete reuse of existing platforms is complex if we enlarge the scope to synergic multimodality (while it is quite “natural” considering a pure mark-up oriented evolution like X+V). For the moment we are following two main roads:

- Using as much as possible standard products and standard technologies. Currently most of the Service Delivery Platform are based on J2EE technologies and, thus, it is quite obvious to move in this direction. We have found that several architectures are based on agents platforms [21]. We consider agent technologies very promising and attractive. However they are not widely accepted: at least in our limited experience, agents platforms are not used for real, proven service delivery platform. Why should they be used for multimodal service delivery platforms? Considering the evolution of the ICT in the last five years, a better approach could be based on web services.
- Using as much as possible open source middleware for creating the conditions of a real diffusion of multimodality. It is a matter of fact that there are very few multimodal services up and running. Probably this is due to the fact that there does not exist a software infrastructure that could be easily used for building new multimodal services. That is why we believe that open source must be taken into account. Thousands of people are able to use open source software, and its quality is now comparable to the best of breed equivalent industrial products. The problem is to select the open source software correctly and find what are the real software modules that must be conceived, designed and implemented ex-novo.
- Design tools/IDE. We have found interesting works presenting new tools for rapid development of multimodal services. These papers are a very good starting point for designing new generation multimodal platforms [20,21,22,23]. We have found especially interesting the attention that the scientific community has dedicated to new languages for describing multimodal interactions. If such a language exists, thus it could be represented graphically and “run” on a specific execution platform. Nevertheless, at the present we have not found any “interpreter” for multimodal languages.
- Client modules: as discussed above, the diffusion of new multimodal services must be seen in the perspective of the real device market. The first priority is to build reusable modules for “composing” new multimodal interfaces. There are interesting works concerning user interfaces that we have found very interesting and inspiring [27,28,29]. More in general, the attention for an abstract representation of user interfaces is growing [27]. It is not surprising that for multimodality XML user interface representation are very good candidates: it is a solution for creating an high personalized, context-aware environment for users. For synergic mobile multimodal services using an abstract representation of the user interface allows to exploit the terminal resources, adding input/output modes depending on the terminal capabilities and, more in general, on the user context. In [28] and [29] the authors demonstrate that it is possible to use XML UI on mobile devices. Their experiments are

always based on high-end mobile devices¹ and, thus, for a small slice of the potential multimodal service market.

Currently we are defining a software architecture, trying to explore possible optimal solution to the issues discussed above. Currently we are selecting and prototyping using as much as possible open source software. This experience could be interesting and reusable and, thus, in the following we discuss what we have found out. We have selected just Java open source, considering the fact that most of the service delivery platforms in the world are based on Java-technologies. In perspective, implementing additional multimodal enablers using Java should simplify the integration phase. The same for the basic libraries for developing mobile clients: considering that our target are java-enabled phones we are considering just J2ME libraries.

Table 1. Open source components for implementing the reference architecture: client modules

Client: open source software		
2D graphics	TinyLine 2D	http://www.tinyline.com/download.html
SVG graphics	TinyLine SVG	http://www.tinyline.com/download.html
XML interfaces	Thinlet	http://www.thinlet.com/
MMI	Piccolo	http://www.cs.umd.edu/hcil/jazz//play/applet/grapheditor.shtml
User Context	OpenDTMP	http://www.opendtmp.org/
UI libraries	J2ME Polish	www.j2mepolish.org/

On the client side, as discussed, the situation is not well established. The Java API are progressing incredibly in the last few years and are usable on the largest range of mobile devices currently conceivable. In Table 1 we report some of the most interesting libraries for building Java-based mobile clients available on internet (apart the ones that are directly installed on the phone from the producer). We believe that they can be used for building a complete framework for creating multimodal interfaces (we are currently working to a first prototype based on them). Our first results are encouraging and this is why we report this experience. *Thinlet* is probably one the most interesting libraries for building XML interfaces. Our idea is to use it for aggregating basic modules (downloaded on the phone over-the-air): as in [29], it will be possible to modify the user interface considering the user context and preferences.

The *Piccolo* library is very well known and used all around the world: it is normally used for graph visualization on fat clients (Personal Computers), but we believe that it can be extended easily for managing multimodal inputs and outputs on mobile devices. We have found that this library could be ported on small devices provided that a library for managing bi-dimensional graphic on small devices exists

¹ In [29] the target language is Java, but it is not specified which Java profile. Since they speak about mobile Java code it is very likely they have used at least the Personal Java profile, since for very limited devices J2ME does not support introspection and, thus, code mobility.

(on the other hand a PocketPC version of the Piccolo framework exists). Considering that, we have found very interesting libraries for managing graphics on small mobile phones, and we are working for building a new enriched version inspired to the Piccolo library. The idea is to have a single environment for managing a graphical representation of the service data and new input-output modes. Thus, we could say that, starting from available open source it is possible to create a complex multimodal environment for creating dynamic, appealing, XML based, multimodal clients. Furthermore, for the java-enabled mobile devices supporting the stylus², we are developing modules for managing handwriting and gesture (including sketches) . Concluding, it is possible to use open source software for building a complete multimodal framework, even if new modules must be designed and implemented, and the basic software packages reported in Table 1 must be extended for supporting a wider range of inputs/outputs.

Table 2. Open source components for implementing the reference architecture: middleware

Middleware : open source software		
J2EE Application server	JBoss A/S	http://labs.jboss.com/portal/
Communication library	JBoss remoting	http://labs.jboss.com/portal/
Session and communication libraries	JBoss Cache/JGroup	http://labs.jboss.com/portal/
Execution environments	Openemcee	http://openemcee.sourceforge.net/
Execution environments	PXE (BPEL) ActiveBPEL (BPEL)	http://www.intalio.com/ www.activebpel.org/
Server localisation	IPTEL SIP server	http://www.iptel.org/
Special SIP services	Cafè SIP	http://www.cafesip.org/projects/jiplet/index.html
Security and session management	JOSSO	http://www.josso.org/
Map Server	GeoServer	http://docs.codehaus.org/display/GEOS/Home

In Table 2 we report some open source that we are using for building our platform. We started from considering that some effort has been spent for the definition of an abstract language for describing multimodal interactions. This works demonstrate that it is possible to give an abstract description of a multimodal interaction. This fact can be reused for developing a graphical design environment for creating new multimodal services. Thus, new services described using graphical building blocks,

² There are very few Java-enabled devices supporting the stylus (QTek and SonyEricsson), even if several new announced smart-phone will support it. For example Nokia will launch a new advanced phone supporting the stylus for the first quarter of the 2007.

are translated in the target multimodal language and executed by an interpreter. Thus, middleware plays an important role: whatever the multimodal language is, it must be interpreted and executed using an appropriate middleware. We are currently using Openmccc, a small project, very simple and not so diffused, but very effective. Among the other features, it does not spawn new threads: thus it is easily embeddable in enterprise objects, while keeping their behaviour consistent.

While other software mentioned in Table 2 are very well known (especially JBoss), probably it must be discussed which kind of middleware must be used for an effective data transmission among mobile clients and servers. Using open source software it is possible to use SIP/SDP for creating a session (and during the handshake establish the communication details) or to use a more traditional approach based on proprietary TCP/IP protocols. It is obvious that a SIP-enabled framework could be more flexible and could exploit all the benefits that such a protocol could offer (for example, it is always possible to know the terminal presence and the user availability). Nevertheless, SIP alone is not enough: a transfer protocol (like RTP) must be implemented. We are currently implementing such a protocol for Java-enabled phone (it exists for Symbian and Windows Mobile) since it is not available as open source.

We have already discussed on a possible integration with pre-existing service delivery platforms. The idea is to use as much as possible web services, eventually orchestrating them, using a BPEL engine (we report a couple of the best).

Table 3. Open source components for implementing the reference architecture: multimodal infrastructure

Multimodal infrastructure : open source software		
Speech Recognition	Sphinx	http://cmusphinx.sourceforge.net/html/cmusphinx.php
Handwriting recognition	Jarnal	http://www.dklevine.com/general/software/tc1000/jarnal.htm
State tracking	Unimod	http://unimod.sourceforge.net/
Rule engine	JBoss Rule	http://labs.jboss.com/portal/
Semantic disambiguation	Wordnet	http://wordnet.princeton.edu/
Context analysis	YALE	http://www-ai.cs.uni-dortmund.de/SOFTWARE/YALE/download.html

We have found open source tools for speech and handwriting recognition. While for speech recognition there are several very good tools (at least for starting up a laboratory), for handwriting the situation is not so mature. Nevertheless the existing tools are good starting points. In other words, an efficient multimodal recognition is still not possible using open source software, but anyway feasible. For a basic fission framework, we are having very good results using open source software. For the service state tracking (sometime called plan management) we are using Unimod. Using this tool it is possible to model the user interaction and the service states as a finite state machine. Unimod offers a very good Eclipse plug-in (www.eclipse.org), that allow to implement, test and deploy a multimodal interaction plans. As discussed previously, our main purpose is to create a complete design suite for multimodal services: Unimod could be a piece of this complete framework.

For a simple input disambiguation the usage of a simple rule engines (e.g. JBoss Rule) and the Princeton WordNet gives good results at least in very simple cases. We are absolutely aware that our work in this moment reproduce well-known techniques for dialog management (our reference is [26], an excellent review of the state-of-art). It could be of some help the fact that some basic tool exists and it is possible to quickly start-up a laboratory using open source software.

4 Conclusions

As discussed in this paper, our work is in progress. We are currently building an experimental proof for validating our architectural approach. The more we study the more we find new issues to face. Our feeling is, however, that our work could stimulate a interesting discussion on multimodal mobile services. This paper, thus, is just a view of the problem from the perspective of a group that comes from different experiences and it is trying to define its own vision of the domain. The issues we are proposing in this paper are related mostly to our direct work experience: the diffusion of mobile devices (and, thus, the potential service market) as well as a comfortable incremental evolution of traditional service delivery platform for supporting synergic multimodality, are issues that, probably, deserve more attention. We hope that this paper could contribute in this direction.

References

1. Bolt, R.A. Put that there: Voice and gesture at the graphics interface. *Computer Graphics*, 1980, 14 (3): 262-270.
2. Oviatt, S.L. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction 12*, (1997), 93-129.
3. Oviatt, S.L., Cohen, P. R., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D., Designing the user interface for multimodal speech and gesture applications: State-of-the-art systems and research directions for 2000 and beyond.
4. Knutkavale, Naradilpwarakagodaand, Janeikesetknudsen Speech Centric Multimodal Interfaces for Mobile Communication Systems . *Elektronikk 2.2003*: 104-117, 2003
5. Garzotto, F.; Mainetti, L.; Paolini, P., 1997, "Designing Modal Hypermedia Applications", The Eighth ACM Conference on Hypertext, Southampton, England, pp. 38-47.
6. Lizhong Wu, Sharon L. Oviatt, Philip R. Cohen Multimodal Integration - A Statistical View, *IEEE Transactions on Multimedia*, vol. 1, n 4, 334-341, 1999.
7. Jennifer L.Beacham, Giuseppe Di Fabbrizio, Mils Klarlund - Towards SMIL as a Foundation for Multimodal, *Multimedia Applications, W3C Multimodal Interaction Activity*, 2002.
8. Oviatt, S. L., DeAngeli, A. & Kuhn, K. Integration and synchronization of input modes during multimodal humancomputer interaction, *Proceedings of the CHI '97 Conference*, New York, NY: ACM Press, 415-422.
9. S. Oviatt, "Mutual disambiguation of recognition errors in a multimodal architecture," in *Proceedings of the Conference on Human Factors in Computing Systems: CHI'99*, pp. 576--583. ACM Press, Pittsburgh, PA, 1999.

10. Oviatt, S. Ten Myths of Multimodal Interaction. *Communications of the ACM*, 42 (11), 74–81, 1999
11. W3C Multimodal Architecture and Interfaces , <http://www.w3.org/TR/2005/WD-mmi-arch-20050422/>
12. EMMA: Extensible MultiModal Annotation Markup Language, <http://www.w3.org/TR/emma/>
13. Overview of SGML Resources, <http://www.w3.org/MarkUp/SGML/>
14. X+V for the Next Generation Web, <http://www.voicexml.org/specs/multimodal/>
15. 3GPP TS 23.057 V4.4.0 (2001-12) 3rd Generation Partnership Project Technical Specification Group Terminals Mobile Station Application Execution Environment (MExE), Functional description, Stage 2 (Release 4) http://www.3gpp.org/ftp/Specs/2001-12/Rel-4/23_series/23057-440.zip
16. Synchronized Multimedia Integration Language (SMIL) Specification, <http://www.w3.org/TR/REC-smil/>
17. Oviatt S. and Cohen. P., 2000. Multimodal Interfaces That Process What Comes Naturally. *Communications of the ACM*, 43(3):45–53.
18. Matera M., SUE : A Systematic Methodology for Evaluating Hypermedia Usability, Ph. D. Thesis, Politecnico of Milano 2000
19. Joyce Y. Chai, Pengyu Hong, Michelle X. Zhou A probabilistic approach to reference resolution in multimodal user interfaces, 2004
20. Elting C., Rapp S., Möhler G., Strube M. Architecture and implementation of multimodal plug and play, Proceedings of the 5th international conference on Multimodal interfaces, 2003
21. Johnston M., Multimodal Language Processing , In Proceedings of International Conference on Spoken Language Processing (ICSLP), Sydney, Australia.1998
22. Filippo F., Krebs A., Marsic I. A framework for rapid development of multimodal interfaces, Proceedings of the 5th international conference on Multimodal interfaces, November 05-07, 2003, Vancouver, British Columbia, Canada
23. Rousseau C., Bellik Y., Vernier F., Multimodal Output Specification/Simulation Platform, Proceedings of the 7th international conference on Multimodal interfaces, Sorrento, Italy
24. Gourdol A., Nigay L., Salber D., Coutaz J. Two Case Studies of Software Architecture for Multimodal Interactive Systems:Voice-Paint and a Voice-enabled Graphical Notebook, In Proceedings of the IFIP WG 2.7 working conference. North-Holland, Aug. 1992.
25. MUST_Multimodal and Multilingual services for small Mobile Terminals. Heidelberg, EURESCOM Brochure Serires, May 2002 (<http://www.eurescom.de/public/projects/P1100-series/P1104/default.asp>)
26. Bui T.H., Multimodal Dialogue Management - State of the art, CTIT Technical Report series No. 06-01, University of Twente (UT), Enschede, The Netherlands, January 2006.
27. Trewin S. Zimmermann G. Vanderheiden G, Abstract User Interface Representations: How Well do they Support Universal Access?, Proceedings of the 2003 Conference on Universal Usability (pp. 77-84). New York: Association for Computing Machinery.
28. Simon R., Wegscheider F., Tolar K. Tool-Supported Single Authoring for Device Independence and Multimodality, in Proc. of the 7th International Conference on Human Computer Interaction with Mobile Devices and Services (Mobile HCI 2005), Salzburg, Austria, September 19-22, 2005
29. Repo P., Rieki J., Middleware Support for Implementing Context-Aware Multimodal User Interfaces, Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia table of contents, College Park, Maryland

MobiSoft: An Agent-Based Middleware for Social-Mobile Applications

Steffen Kern¹, Peter Braun², and Wilhelm Rossak¹

¹ Friedrich Schiller University Jena, Computer Science Department
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{kern, rossak}@informatik.uni-jena.de

² The agent factory GmbH
Leutragraben 1, 07743 Jena, Germany
braun@the-agent-factory.de

Abstract. We present an agent-based middleware for social-mobile applications, which has been developed as part of an ongoing linkage project. One aspect of the *MobiSoft* project is the vision of facilitating, augmenting, and promoting human social interaction by electronic personal assistants during face-to-face encounters. Possible areas of social mobile applications include the establishment of groups or communities based on shared interests or goals, the exchange of information such as personal profiles, news, private sales, or any kind of recommendations, and the preselection of possible communication partners in social networks. We outline the decentralized peer-to-peer based architecture and present techniques for information representation using semantically rich languages based on existing standards. We describe, how mobile agents are facilitated as user representatives and intelligent information carriers in mobile ad-hoc networks and present a first prototype of a social-mobile application.

1 Introduction

In the last few years, mobile devices, in particular mobile phones, have become part of our daily life. They are already our indispensable companions and are powerful enough to provide services beyond plain phone functionality, for example to help us managing appointments, contact lists, and personal tasks.

We expect mobile phones to be an important means in future for setting up and maintaining social networks between people. First approaches have already been presented, for example Dodgeball is an SMS-based notification service for friends and acquaintances in proximity. An approach for localization and immediate communication between members of a group is presented in [1]. Beyond these first approaches for network maintenance, we envision new application scenarios for *creating* social networks taking the opportunities of personal area networks into account.

Our *MobiSoft* project is driven by the vision of facilitating, augmenting, and promoting human social interaction by electronic personal assistants during face-to-face encounters. The main idea is to let autonomous software agents act as

personal representatives and search for possible communication partners in the *digital space* that is created by a personal area network based on Bluetooth or WiFi around mobile phones. Agents can communicate with each other, for example for exchanging personal profiles and task lists. Finally, a pair of agents can notify their respective owner for continuing this process in the real world. In the following, we call these agents *social-mobile assistants*.

Areas of such social-mobile assistants include the establishment of groups or communities based on shared interests (work, hobbies) or activities and goals (such as to reduce travel costs by sharing a taxi). Social-mobile assistants can exchange information such as personal profiles, news, private sales, and preselect possible communication partners in social networks. It will be possible for them to coordinate shared task lists and diaries by automated negotiations.

In this paper we will describe our approach in detail, in particular the underlying agent technology, and the current state of our implementation. The rest of this paper is structured as follows: The next section discusses similarities and differences to other projects and Section 3 outlines the goals of our research project. The following section describes the architecture of our approach and the current state of implementation. Finally, the last section gives an outlook to future development.

2 Related Work

Current approaches for social-mobile applications [2] are mostly based on central servers and text messages. For example, Dodgeball and Playtxt are social-mobile networks to locate friends, friends of mutual acquaintances or other people with matching profiles. In those applications a user has to provide his or her current location manually, whereas in the Reno system [2] the current location is determined via GSM technology. Our approach differs from these techniques in that we focus on personal area networks, which can be seen as a digital space around a person, whose size depends on the underlying wireless transmission technique. If two digital spaces overlap, people can virtually see each other, that is, their mobile devices are able to exchange information. In a personal area network we do not need an explicit notion of places and provision of proximity information is an inherent network function.

Looking into research in the area of mobile ad-hoc networks, we see that most research has focused on the problem of multi-hop routing data packets to enable Internet-like applications in ad-hoc networks. In particular, they address the issues of how to enable peer-to-peer like applications on mobile devices [3] to share files [4,5], MP3 play-lists [6], or information dissemination of homogeneous data of one application domain such as traffic information [7]. Most of those approaches were developed mainly for enabling information exchange that is manually triggered by users [6] or make information dissemination completely independent of the user [7]. In contrast, our project aims at the establishment of social interactions by use of mobile ad-hoc networks and tries to overcome the application-specific data modeling and restriction on specific application

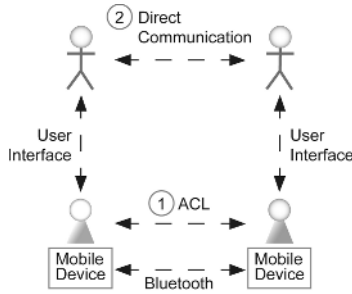


Fig. 1. Interaction between agents running on mobile devices and their users in our approach. At first, agents start to communicate together before in a second step, agents delegate responsibility for further communication back to their users.

domains. In [8] we have first introduced MobiSoft to a wider audience and described the projects aims as well as early ideas and concepts.

3 The MobiSoft Project

The MobiSoft project is an ongoing linkage project including Friedrich Schiller University Jena, the agent factory GmbH, and Godyo AG and is funded by the Thuringian Ministry of Economy, Technology and Labor in Germany.

In MobiSoft we aim at application scenarios, in which humans travel around and meet at specific places, for example shopping malls, sports stadiums, public transport, museums, libraries, conferences, lecture halls, etc. Although it might be helpful and interesting, people rarely start talking to complete strangers, because of inhibitions, social barriers or simply a lack of time. Otherwise, if people knew each other, they would more freely exchange information and, therefore, spread and receive pieces of useful information that could further be combined with already existing information and forwarded to others.

In this project we aim at supporting such a very human behavior of information exchange with software agents [9] as user representatives that reside on mobile devices and to which users have delegated the task of finding proper human communication partners, compare Fig. 1. We conceive software agents to be small entities that are situated in a networked environment of mobile devices. Agents are able to react on percepts from the environment about other agents in proximity and then autonomously commence information exchange with them. Communication between agents is based on messages, which are annotated with semantic information as defined in an *agent communication language* using high-level communication protocols such as negotiations [10].

This information exchange works transparently for the user only in the first step, in which the agents exchange information, such as user profiles, or negotiate best interaction time. Later, agents inform their respective users about the potential communication partner and let them decide on further steps. By this,

we overcome existing inhibitory behavior of humans by delegating this task to software agents, while the agents' goal is to find *proper* communication partners and *interesting* information. The project focuses on developing a new framework for social peer-to-peer information exchange in mobile ad-hoc networks. It has the following key aims:

- Develop efficient agent-based techniques for managing peer-to-peer overlays in mobile ad-hoc networks. In particular, we restrict ourselves to Bluetooth, since most of today's mobile phones support this network technique rather than WiFi.
- Develop hybrid information exchange techniques, in which mobile agents pro actively distribute information to as well as reactively receive interesting information from other agents, taking into account the specific limitations of mobile ad-hoc networks.
- Develop methods to describe user profiles, interests and information using semantically rich languages, which are based on existing standards known from the Semantic Web. Efficient techniques have to be developed for matching user profiles while taking into account the specific hardware limitations of mobile devices.

We are aware of several additional research issues, for example in the area of privacy protection and human-computer interaction to make this type of application both useful and acceptable by users. We see this project as a first step in which we aim at developing the framework and technical infrastructure that will also enable later studies of those issues in detail.

4 Architecture and Implementation Details

The MobiSoft middleware consists of a multi-agent system that is distributed over mobile devices. We use the Java-based Tracy2 agent toolkit [11] as foundation for our middleware, because of its clear separation between basic and extended (optional) functionality, which enables a straightforward incremental down-sizing process while taking into account the restrictions of mobile devices.

The MobiSoft middleware is divided into three layers, where the lowest layer contains the micro kernel of the Tracy2 agent system. It provides basic functionality for thread scheduling and maintaining the life-cycle of software agents. The middle layer of our architecture is formed by several plugins for Tracy2. Some of them were taken from Tracy2 without any or only slight modifications, while others were newly developed to match the specific functional requirements of social-mobile assistants. Finally, the top layer of our architecture contains the social-mobile assistants, which are mobile software agents that provide the business logic of the whole application. The lower two layers of our architecture have to be deployed on all mobile devices upfront, whereas agents are deployed at run-time at the earliest and can easily be replaced in case of newer versions.

Based on the techniques described below we developed a prototype running on Bluetooth-enabled mobile phones for this year's CeBIT exhibition. This

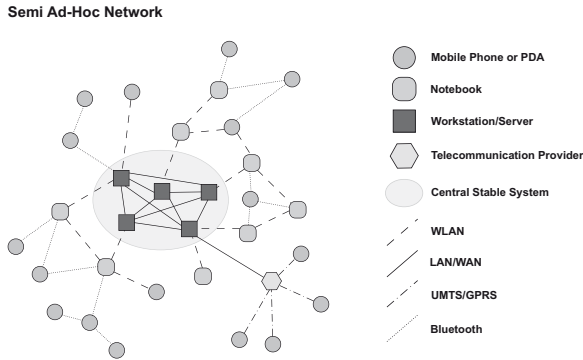


Fig. 2. Structure of a Semi Ad-Hoc Network

prototype allowed business people to find communication partners with similar interests using peer-to-peer Bluetooth communication and let their personal assistants negotiate on a suitable appointment. User interests were defined by browsing an exhibition catalog. We will outline our experience in the last part of this section.

4.1 Peer-to-Peer Network Overlay Plugin Using JXTA

As mentioned above, we aim at application scenarios where mobile users meet each other and establish ad-hoc networks. Despite of social-mobile applications, which can be implemented solely on networks established between mobile devices, MobiSoft does not disregard application scenarios which require hard-wired powerful backbones that consist of a number of workstations. For example, this kind of network structure can be found in a shopping mall, in which shops or restaurants establish a network of stationary workstations providing Web services for product information or restaurant bookings. We call those hybrid networks *semi ad-hoc* and Fig. 2 depicts that case.

To structure semi ad-hoc networks we will apply a peer-to-peer approach using JXTA [12]. We decided on JXTA, because it is available for various platforms including J2SE and J2ME environments and powerful peer-to-peer routing mechanism for mobile devices in ad-hoc networks base on JXTA have already been proposed in literature, for example [13]. We employ JXTA in two use-cases:

- To publish and find information about available Tracy2 platforms. This includes the discovery of mobile devices in proximity.
- To publish and find information about available services. This allows agents to find desired functionality or information dynamically and independent of a specific platform.

Nevertheless, the choice for JXTA should not lead to a tight coupling between Tracy2 and JXTA, thus we aimed at providing an abstract network management

which is just a wrapper for any type of network overlay. The plugin provides a fixed interface for other plugins and agents to find Tracy platforms and search for respectively publish services. It will delegate those requests to a network overlay module – in our case JXTA – which is capable to provide those services. Using such a modular plugin architecture, it would be easy to replace JXTA with one or more better alternatives without changing the plugins interface. That would have led to changes in all agents and plugins that use the network management.

4.2 Mobile Agents on Mobile Devices

On top of the network overlay technique and some other basic plugins of Tracy2, which we cannot mention here due to a lack of space, mobile software agents [14] are used for providing the whole business logic of social-mobile assistants.

The reason to employ mobile agents in this project is simply its beneficial design paradigm compared with other more traditional paradigms such as client/server. Using mobile agents, it is not necessary to define complex network protocols for the transmission of information units because mobile agents carry the protocol. Modifications of the protocol do not imply changes to the two lower layers of our architecture and, therefore, it is not necessary to deploy the application to mobile devices again. With this approach we are able to implement existing strategies for information dissemination and are open for future enhancements. In addition, mobile agents have been proven theoretically to work very well in mobile environments and to be in particular robust against network failures.

Although, the benefits of mobile agents on mobile devices are already known for a couple of years, no Java-based mobile agent toolkit is available on mobile devices, yet. The reason for this is that Java's mobile edition does not support many of the features necessary for mobile agents, such as object serialization, class loaders, and dynamic code downloading. Our approach for enabling mobile agents on mobile devices makes use of a virtual machine for a proprietary agent-oriented programming language. This virtual machine defines control flow statements such as conditionals and loops, and a basic set of commands necessary for agents to access high-level functions provided by the agent toolkit and its plugins. We have developed an own internal representation based on abstract syntax trees [15], which are stored as data objects in an agent's data store. By doing so, we have chosen the obvious solution to the problem of code transmission by translating code into data. Each agent uses it's own virtual machine for this programming language. Data serialization is also implemented straightforward by own mechanisms to flatten objects recursively into byte streams.

4.3 Mobile Agents as Information Carriers

Mobile agents are injected into the system and then roam autonomously from peer to peer to distribute the information they carry as their data. Mobile agents are aware of their environment, that is, devices and other agents in their vicinity, and

communicate spontaneously to other agents. After careful consideration of different strategies for information dissemination [16] such as *flooding* or *broker-based* approaches [17], we have decided on a technique known as *epidemic dissemination*, where a mobile agent carries an information unit to a randomly chosen group of mobile nodes. This dissemination approach enables messages to propagate quickly in the network and it is very robust against the node and network link failures. For more information on those algorithms we refer to [18,19]. So far, the epidemic-based algorithms have only been studied as a general replacement for traditional routing and multicast algorithms in mobile ad-hoc networks. The mobile agents' code can be considered to define the dissemination strategy, for example the scope of the information unit, that is, the distance from the information source, the spatial direction and temporal freshness [20].

Our approach spreads information units in two ways. First, they virtually hop in a form of mobile agents from device to device. Second, as the users move physically through an environment, they distribute information to new areas.

4.4 Stationary Agents as User Representatives

Stationary agents are used on mobile devices as user representatives. They act as data consumers as well as data providers, to protect users' private profile and to protect users against information spamming. Stationary agents communicate to arriving mobile agents and launch mobile agent themselves to explore the environment for other agents. Since we are working in open environments, which should not be restricted to one specific application domain, we aim at a flexible and extensible approach for matchmaking between user profiles and interests. Although, there has been a lot of research done in this area already [21], today's available techniques are still quite limited.

For example, *www.tribe.net* is a Web site enabling people to find other people based on their interest, given by keywords. On mobile devices, we find *www.upoc.com* or *www.jambo.org* to establish communities and locate each other based on keyword-based profile matching. Neither approach uses semantic descriptions of user profiles and preferences, but only simple text-based approaches.

Friend of a Friend Finder (FOAF) is a project that aims to share information about persons in the Internet. The language used in this project is RDF that provides a means to describe data and meta-data. RDF defines a simple data model which consists of resources and statements that link two resources, comparable to a subject-verb-object relationship. A statement is called a triple, which consists of subject and object, and the predicate that plays a role of the verb mentioned previously. In a so-called FOAF file, a user describes his personal data and which other persons this user knows. With a help of these links to other persons, a search engine can now create a graph of who knows whom. However it is not possible to describe interests and preferences with FOAF. We have adapted FOAF profiles and added appropriate information to describe interests. Figure 3 shows an example RDF profile.

Whenever user agents receive new information from roaming mobile agents, they have to match it to user interests on a semantic level. New information will

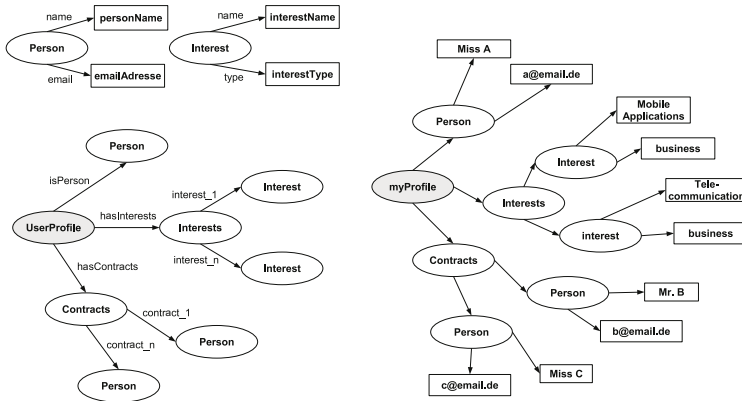


Fig. 3. User Profile Description in RDF. On the left hand side, the structure of a user profile in RDF is given. On the right hand side, one can see a concrete instance of such a profile. For the sake of clarity we have omitted Uniform Resource Identifiers and skipped all arrow labels on the right hand side.

be presented to the user only if it is important and related to the current context. Otherwise, it will be inserted to a pool of pending information units. The agents will improve their internal user model by learning from user feedbacks about the relative importance of information. Other information units, not important to the user at the moment, will be tolerated to some extent to support the general dissemination strategy of our approach. The user agents also sort out out-dated or falsified information from time to time, if necessary. Access to the user profiles is protected by the user agents. Whenever a mobile agent gains access to sensitive profile information, the user is requested for acknowledgment. Again, these user agents are able to learn users' intentions and in the longer term act autonomously on behalf of their owner and provide access to the user profile in a very fine-grained way.

4.5 Lessons Learned

As mentioned in the introduction of this section, we developed a prototype based on this architecture for the CeBIT exhibition. While running stable and fast in a controlled environment, the application proved to be nearly useless at the exhibition due to Bluetooth limitations. Even when lying right next to each other, our mobile phones were, in most cases, unable to detect each other. They rather found some of the numerous Bluetooth-enabled devices like notebooks, keyboards or headsets that crowded the exhibition. The problem not only occurred in our application, even on operating system level we were unable to establish a link between two devices. This is unfortunate, but we expect that upcoming device generations can better fit our needs.

5 Conclusions and Outlook

The aim of MobiSoft is to deliver a new architecture for social-mobile applications which is based on existing techniques like mobile agents, peer-to-peer networking and semantic descriptions of services and profiles. After years of in-depth research on mobile agents and mobile applications we are confident that such a combination can lead to an adaptable and powerful middleware.

We will now move on to improve user profile and service descriptions and tackle open issues like service propagation, dynamic service binding and profile matching. One of our next steps will be to integrate several semantic matching algorithms/systems into Tracy and evaluate and compare their suitability.

Motivated by the experiences with the CeBIT prototype, we will establish a test bed at the university campus. This test bed will provide university information, community functions, and access to the library system for students and employees all around the campus. The users feedback, both on hard- and software issues, should help us to improve the whole system. Furthermore, we will be able to conduct comprehensive performance and reliability experiments.

Acknowledgments

The work presented in this paper is partially funded by the Thuringian Ministry of Economy, Technology and Labor under grant FKZ B 509-04005.

References

1. Smith, I.E., Consolvo, S., LaMarca, A., Hightower, J., Scott, J., Sohn, T., Hughes, J., Iachello, G., Abowd, G.D.: Social disclosure of place: From location technology to communication practices. In Gellersen, H.W., Want, R., Schmidt, A., eds.: *Pervasive Computing, Third International Conference, Munich (Germany), May 2005*. Volume 3468 of *Lecture Notes in Computer Science.*, Springer Verlag (2005) 134–151
2. Smith, I.: Social-mobile applications. *Computer* **38**(4) (2005) 84–85
3. Oberender, J., Andersen, F.U., de Meer, H., Dedinski, I., Hossfeld, T., Kappler, C., Maeder, A., Tutschku, K.: Enabling mobile peer-to-peer networking. In Kotsis, G., Spaniol, O., eds.: *Mobile and Wireless Systems*. Volume 3427 of *Lecture Notes in Computer Science.*, Springer Verlag (2005) 219–234
4. Christoph Lindemann, O.P.W.: A distributed search service for peer-to-peer file sharing in mobile applications. In: *Proceedings fo the Second International Conference on Peer-to-Peer Computing (P2P02)*, IEEE Computer Society Press (2002)
5. Gang Ding, B.B.: Peer-to-peer file-sharing over mobile ad-hoc networks. In: *Proceedings fo the 2nd IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMM04)*, IEEE Computer Society Press (2004)
6. Wiberg, M.: Folkmusic a mobile peer-to-peer entertainment system. In: *Proceedings of the 37th Hawaii International Conference on System Sciences*, IEEE Computer Society Press (2004)

7. Ouri Wolfson, Bo Xu, A.P.S.: An economic model for resource exchange in mobile peer-to-peer networks. In: Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM04), IEEE Computer Society Press (2004)
8. Kern, S., Braun, P., Dettborn, T., Eckhaus, R., Ji, Y., Erfurth, C., Rossak, W.: A generic agent-based peer-to-peer infrastructure for social-mobile applications. In Kirste, T., Knig-Ries, B., Pousttchi, K., Turowski, K., eds.: Mobile Informationssysteme - Potentiale, Hinternisse, Einsatz, 1. Fachtagung Mobilitt und Mobile Informationssysteme (MMS 2006), Passau (Germany), February 2006. Volume P-76 of Lecture Notes in Informatics., Springer Verlag (2006) 127–138
9. Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley and Sons (2002)
10. Weiss, G., ed.: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence. MIT Press (2000)
11. Braun, P., Müller, I., Schlegel, T., Kern, S., Schau, V., Rossak, W.: Tracy: An extensible plugin-oriented software architecture for mobile agent toolkits. In Calisti, M., Klusch, M., Unland, R., eds.: Software Agent-Based Applications, Platforms and Development Kits. Whitestein Series in Software Agent Technologies. Birkhäuser Verlag (2005) 357–382
12. JXTA (<http://www.jxta.org>)
13. Buccafurri, F., Lax, G.: Tls: A tree-based dht lookup service for highly dynamic networks. In: CoopIS/DOA/ODBASE 2004, LNCS 3290, Springer Verlag (2004) 563–580
14. Braun, P., Rossak, W.R.: Mobile Agents—Basic Concept, Mobility Models, and the Tracy Toolkit. Morgan Kaufmann Publishers (2005)
15. Aho, A.V., Sethi, R., Ullman, J.D.: Compilers: Principles, Techniques, and Tools. Addison-Wesley (1986)
16. Franklin, M.J., Zdonik, S.B.: Dissemination-based information systems. Data Engineering Bulletin **19**(3) (1996) 20–30
17. Eiko Yoneki, J.B.: An adaptive approach to content-based subscription in mobile ad-hoc networks. In: Proceedings of the 2nd IEEE Annual Conference on Pervasive Computing and Communication Workshops (PERCOMM04), IEEE Computer Society Press (2004)
18. Kermarrec, A.M., Massoulie, L., Ganesh, A.J.: Probabilistic reliable dissemination in large-scale systems. IEEE Transaction on Parallel and Distributed Systems **14**(3) (2003) 248–258
19. Vogels, W., van Renesse, R., Birman, K.: The power of epidemics: robust communication for large-scale distributed systems. ACM SIGCOMM Computer Communication Review **33**(1) (2003) 131–135
20. Marco Mamei, F.Z.: Programming pervasive and mobile computing applications with the tota middleware. In: Proceedings of the 2nd IEEE International Conference on Pervasive Computing and Communications. Orlando, FL (USA), March 2004, IEEE Computer Society Press (2004)
21. Kleemann, T., Sinner, A., von Hessling, A.: Semantic user profiles and their applications in a mobile environment. In: Workshop on Artificial Intelligence in Mobile Systems at UbiComp 2004, Nottingham (UK), September 2004. (2004)

Innovative Healthcare Services for Nomadic Users

Marcello Melgara¹, Luigi Romano¹, Fabio Rocca¹, Alberto Sanna²,
Daniela Marino², and Riccardo Serafin²

¹ Atos Origin Italia S.p.A., 11026 Pont-Saint-Martin (AO), Italy

{marcello.melgara, luigi.romano, fabio.rocca}@atosorigin.com

² San Raffaele Scientific Institute, e-Services for Life and Health, 20132 Milano, Italy

{alberto.sanna, daniela.marino, riccardo.serafin}@hsr.it

Abstract. Mobile users require location based, situation aware services, especially when healthcare is involved. Within “Nomadic Media” an Eureka-ITEA International projects, Ontology based semantic web service discovery and orchestration have been studied and applied to provide mobile users with innovative healthcare services.

The designed system identifies, orchestrates and customises the services, according to the health status of the user and his usage location, usage conditions and environmental situations.

The paper will describe the developed technologies and the implemented services.

Similar approach was followed in PIPS: “Personalised Information Platform for Life & Health Services”. However in PIPS more emphasis has been applied in the definition and the development of Use Cases and service definition and development for Patients and Citizen.

1 Introduction

The PIPS Project¹ is an ongoing Integrated Project within the IST 6th Framework Programme, coordinated by Fondazione Centro San Raffaele del Monte Tabor. Together with FCSR and Atos Origin, other 15 Companies and Universities from Europe, Israel, Canada and China cooperate to it. The PIPS project objective is to encompass the entire set of business processes, professional practices, and products, applied to the analysis and preservation of the citizen’s well-being, using the latest innovations in ICT. The project joins healthcare (HC) suppliers, citizens, public organizations, food/drug industry and services, researchers, and health related policy makers. These actors create a dynamic knowledge environment that feeds the system and gives added value feedback for personalized contextual knowledge and services to improve the European public’s wellbeing. In the PIPS context each actor is a supplier and receiver of personalized knowledge. This includes both explicit and implicit knowledge management based on

¹ IST 2004 507019 PIPS: Personalised Information Platform for Life & Health Services, www.pips.eu.org.

traditional and new approaches to knowledge discovery out from current medical practice, evidence based medicine, and disparate knowledge sources from health/nutrition domains.

The Nomadic Media² consortium was composed by different European companies such as Nokia, VTT, Philips Digital Systems, Cefriel and Atos Origin. The Nomadic Media Project aimed to enhance consumer flexibility in the use of services and contents in the places they wish, and to enable the movement of content between their preferred devices according to their needs and circumstances. Implicit in this vision is also the need for consumers to configure services and content in the ways that suit their particular circumstances and thus enjoy the benefits of a personalized environment.

2 eHealth Services

In the following chapters we describe the approaches in the multimodal and pervasive eHealth services adopted in the two research projects Nomadic Media and PIPS.

2.1 The PIPS Service Oriented Approach

PIPS implemented solutions enable:

- HC Professionals to deliver just-in-time personalized and prevention-focused HC services compliant with the Citizen's personal health state, preferences and ambient conditions
- Citizens to make informed decisions about therapies and nutrition at any time and place according to the real-time evaluation of their health state
- HC Authorities to improve risk management of HC systems

The current PIPS project implementation foresees three scenarios: Diabetes, Hearth Failure, and Nutritional. Figure 1-A depicts an example of possible functionalities that could be implemented for the scenarios.

In order to better explain which kind of services the Platforms allows, we are going to present one fictional scenario, completely supported by the current version of the PIPS Platform. In the remaining of this section we will then present the technical architecture employed by the Project, which allows to implement such scenario.

Our main actor for this scenario will be John Fitzgerald, a 55 year old person who is currently under treatment for an ischemic cardiomyopathy with heart failure (HF) complication. This type of disease usually impairs the functional capacity and quality of life of affected individuals.

For this kind of patient it is mandatory to monitor vital signs and the arise of symptoms indicating possible disease accentuation [1]. For this reasons, John is following a complex treatment made of different components: he has to take a

² Eureka E!2023 - ITEA if02019: Nomadic Media: Entertainment at home and on leave. ITEA - Information Technology for European Advancement, <http://www.itea-office.org>

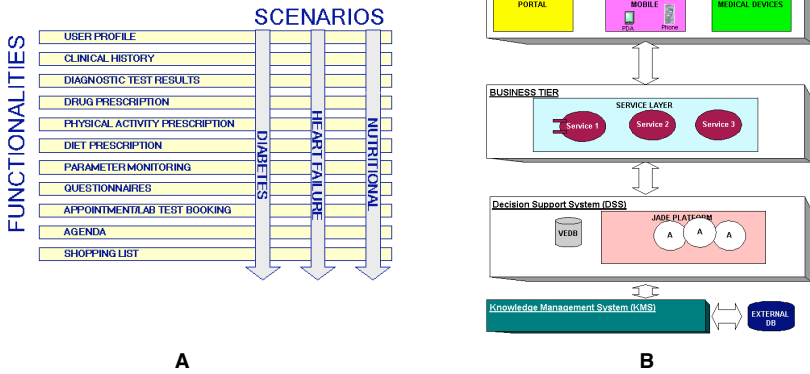


Fig. 1. PIPS Scenarios and possible services (A) and PIPS overall architecture (B)

quite large number of pills daily, he must measure a fixed set of vital signs every morning and he is on a diet.

Therefore, the first support that PIPS provides to John is a reminder service, set with respect to all his prescriptions and integrated with his mobile phone agenda, that reminds him of all actions he must perform on a timely basis.

In John’s case, he must measure his weight, blood pressure and heart rate every day before breakfast. Monitoring these signals daily is fundamental to keep his overall health status controlled and to react properly to any abnormal condition, for instance to accumulation of the body fluids.

For this scenario let’s suppose that John was out of town for the weekend, exceeded a little with drinking and forgot to take some of his pills. When he receives the reminder on Monday morning he uses the measurement devices provided by PIPS and collects the needed vital signs. These devices are all connected wirelessly to John’s home network and the data collected are immediately sent to the PIPS system via a web service.

In John’s case, the analysis of the collected vital signs (increased weight, low blood pressure, high heart rate) indicates a body fluids accumulation, which is confirmed by the symptoms collected using an online questionnaire (weakness, shortness of breath, reduction of diuresis, etc.) and additional vital signs (low oximetry). As a result, he is suggested to resume his diet (strictly adhering to his water intake regime) and a new therapy with an augmented dosage of diuretic is set by the system for him (updating his agenda). The diuretic therapy change is prescribed beforehand by John’s doctor who, along with the normal dosage, will provide the augmented dosage to be taken in case of suspected fluid accumulation.

John’s doctor is also warned of the situation. On her PIPS Portal home page, or cellphone if so configured, she will receive a message that John’s therapy has been changed according to her indication and she will be able to check the collected data (both vital signs and questionnaire answers). At that point she can take an informed decision and react appropriately (for example, fixing John’s therapy, contacting and reassuring him or inviting him for a checkup).

This scenario demonstrates some of the key benefits that a system like PIPS can provide: constant monitoring of patient health status enabled by the integration of self-care networked measuring devices, complex decision support system and new communication technologies, just in time abnormal health condition detection for chronic patients, integrated access to patients continuity of care records, personalized advices to improve health status. Moreover, all these functionalities are personalized according to the context, the situation and the adopted communication devices.

The architecture applied for PIPS system is based on different tiers as shown in Figure 1-B: Front End, Business, Decision Support System, Knowledge Management.

- Front End Tier: it is the collector of the requests incoming from users that can interact with PIPS system through the Portal web application, personal mobile devices and medical devices.
- Business Tier: it implements the Business Logic of the PIPS system, interacting with Front End tier and Decision Support System
- Decision Support System (DSS): it is the technological core of the delivered PIPS system and is based on multi agent platform [2]. DSS processes the information incoming from users in order to monitor the Patient health and wellness status and provide suggestions that should be performed by the PIPS actors [3].
- The Knowledge Management (KM): it provides Knowledge Base information that can be used by DSS and by the final user in order to retrieves trusted e-health information [4].

The DSS is the technological core of the delivered PIPS system and is based on multi agents platform and rule engine. The analysis and design of the system was based on the notion of *computational organisations*, whereby the roles played by the agents are modeled on the basis of the roles in real-world organisations. For the high-level analysis and much of the design was guided using the Gaia methodology [5], in which roles are a central concept.

The model is based on real-world health-care systems. Using the Gaia methodology, two main roles within the PIPS DSS were derived, personal and specialized agents, as follows: citizens have personal advisory agents, what in many organisations are “personal assistants”, which provide personal information to them, such as their diary, or their medical history; specialised agents represent the health care specialists and nutritional experts that assess the health and diet of the citizen and provide advice or information for the nutritional and medical fields in which the PIPS system specialises, such as diabetes or heart problems.

2.2 The Nomadic Media Service Approach

To reach its aim the Nomadic Media project explored different usage scenarios: “At the Airport”, “On-the-Go”, “At Home” and “Healthcare”. The Healthcare scenario was the one in which Web Services (WS) [6] were investigated. For this scenario a technological architecture was defined to be able to:

- Connect a variety of related services into a coherent set and thereby improve the process of, for example, ordering prescriptions, making patient appointments, and scheduling laboratory tests by healthcare professionals
- Enhance physician productivity with real-time access to information via a variety of preferred devices, adapted to the variable usage condition (situational awareness)
- Allow collaborative access for different users such as insurance companies, healthcare providers, drug companies and patients.

As a result of this investigation we realized that the healthcare context shares many characteristics that are common to many complex and distributed applications. This led to the following summary of the key problems:

- Services should be composed at runtime, based on a multiparty business process model, using an orchestration framework.
- To enhance the choices and to dynamically compose the services, advanced techniques should be used to advertise and discover them.
- Content and level of services provided should be adapted in relation to context, situation and user preferences.

A Service-oriented Architecture (SOA) is a possible way to solve the above mentioned problems, that is to say an architecture where functionalities are implemented, essentially, as a collection of services communicating with each other. A service is a function that is well-defined, self-contained, and does not depend on the context or state of other services. As known, SOA is not new, but is an alternative model to the more traditionally tightly-coupled object-oriented models that have emerged in the past decades. Web Services (WS) represent a set of specifications defining the details needed to implement services and interact with them. Although WS is a technology in development and standardization efforts are not completed, robust enterprise toolsets are available: industrial WS based solutions can be developed in specialized areas. One of the main areas explored in Nomadic Media was service composition. It allows complex tasks to be executed as sequences of processes written using standard specifications.

In the Nomadic Media Healthcare scenario the need to exchange information between different providers (services, content and context providers) was clearly identified. WS standards were used to solve the problem of obtaining robust multiparty interaction. Standards also helped to solve the interoperability problems at a syntactic level; however, the real strength of WS technologies is the possibility to afford the heterogeneity of the systems in a semantic way too. To achieve semantic interoperability, information systems must be able to exchange data in a way that allows ready accessibility to the precise meaning of the data and the data itself can be translated by any system into a form that it understands. This was achieved by describing, functionally and operationally, services in a formal, machine-readable way.

Semantic interoperability enables the automation of some procedures:

- Web Service discovery; the action of matching available service descriptions to a requester's candidate service query and returning the resulting matches,

- Web Service invocation; the principle of automatically interacting with an atomic service by using the semantic description to understand how to access it,
- Web Service selection and composition; the action of choosing the most suitable service(s) from a set of known services and running a composite service by invoking WS, in the correct order, overcoming syntactic, structural, semantic and process heterogeneity, and handling errors and exceptions,
- Web Service execution monitoring; the principle of tracking what is happening to some described aspects of a service and its component services.

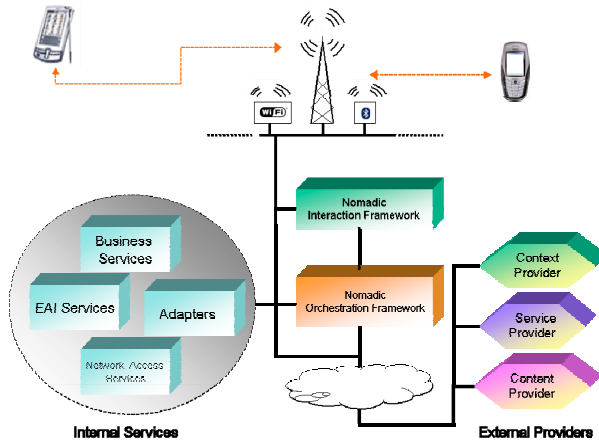


Fig. 2. Nomadic Media Structure

One possible way to proceed toward the semantic interoperability is to provide semantics using metadata and ontologies.

The two major efforts in defining a Semantic Web Services Language [7] are SWSL developed by the OWL-S (Web Ontology Language for Services) [8] committee in the USA, and WSML (Web Service Modeling Language) [9] developed by the Web Service Modeling Ontology [10] project in the EU. A table presented in [11] summarizes the comparison between WSMO and OWL-S.

However the approach to semantic web services does not only engage the language problem, but also the architecture. Since the WSMO project also proposes a framework (WSMF) to solve semantic interoperability, we chose this approach in our project. Its philosophy is based on two (complementary) principles; maximum de-coupling of its components, and a scalable mediation service.

One of the important points we found in the WSMO was the possibility of enabling support for static and dynamic composition. The composition approach we chose was defined as Orchestration. Orchestrator is the one who defines how the overall functionality is achieved by the cooperation of more elementary services. In the orchestration approach a workflow process invokes a number of

different services in a specific order because they have data and control dependencies between each other. Infrastructure based on this approach just requires one central service that is the workflow engine that controls and executes the entire workflow process. The engine we used executes an XML-based [12] script language: BPEL4WS [13]. BPEL4WS (Business Process Execution Language for Web Services) provides a language for the formal specification of business processes and business interaction protocols. By doing so it extends the WS interaction model and enables it to support business transactions.

2.3 Multimodality and Content Adaptation

Multimodality is the technology whose purpose is to enhance the user experience by enabling service providers to combine different ways to build-up intuitive and powerful applications.

For users, multimodality represents an efficient way to interact with a device. For network operators, the combination of audible and visual functions represents the future of communications. Soon, applications such as the mobile ones, will take advantage of multiple simultaneous channels of communication, leading to a new wave of service offerings. Some new standards are focused on multimodality technology [14] [15].

Multimodality and adaptation of contents cover an important role in Health services. Multimodality allows the possibility to transfer information through different channels, while adaptation can be read according to a more general meaning, depending not only on the features requested by the device (hardware and software), but also on particular contexts in which the request is made. A generic eHealth framework should manage situations in which the user cannot receive complete information, but just an abstract of them, or he would like to temporally freeze and save this information.

In particular, when user's presence is revealed, his state is communicated to the system, together with his preferences and his possibility to make use of nomadic services; for example it is possible that a user would not like to be disturbed by unexpected communications.

Besides, the user physical collocation makes easier the service interaction: information related to the structure in which the user is located, are immediately provided (i.e. pharmacies, hospitals) so he can be better assisted in his choices.

The PIPS system offers to final users different access mode in order to acquire information data, related to vital signs user data, food and drug information.

The user can provide data to the system manually, through web Browser UI, or mobile application (e.g. transcript the vital signs values from traditional device to the web form provided by system). The same type of data can be directly sent to PIPS system if are used wireless medical devices connected to the custom applications installed in the mobile devices or PC.

The user can send to the system food and drug information data interacting in different mode: optical mode, RFID technology. For example using the camera of user mobile phone, the picture of the product bar code is processed by an OCR, in order to send the food information data directly to the PIPS system. The

same information could be sent to the system by filling the web form provided by the mobile application with the barcode number.

Also the RFID technology is used if the food or drug product data are store in tag RFID.

The user can also provide inputs to the system, writing on a paper, using an optical pen connected via Bluetooth to the mobile device. Also in this case the data can be provided automatically to PIPS system.

In Nomadic Media the solution approach is based on an adaptation engine service: it uses the client features to dynamically support the device context.

The page rendering is performed at run time, whenever possible, depending on wireless device multimedia capabilities. The Nomadic Media Framework, using database devices service [16], is able to display images and adapted content. It supports, also, Voice Interaction [17] used for manage healthcare vocal application.

3 Conclusions

The technologies studied and applied in PIPS and Nomadic Media have shown their powerfulness to build Healthcare and Wellness Services for Citizen and Patients. This technological enablers have allowed to define innovative complex Service models, derived and sustained by a strong Business Strategy.

In order to make the services more and more intelligent and automated, machine processable data have been adopted, to deal with dynamic content and services. In order to enable the interoperability among systems, a framework using machine processable data, rather than more human oriented data, should be developed. Process ontologies, rule based Multi-Agent environments, and ontology bridging are the most promising avenues for integrating automation and orchestration.

According to the practical experience, the technology required to make eHealth services more “intelligent” (meaning that they are tailored on user) should be explored based on the work in the field of semantic web and multi agent platform. Taking into account the different environments in which eHealth and Wellness solutions may be exploited, the Service should automatically adapt their behavior to the current environment (Situational Awareness) and allow to use the most suitable communication channels and mode (Multi-Channel Multi-Modal Interaction). The aforementioned development lines should allow to concretely achieve the condition that Tim Berners-Lee described as the “Next Generation Web” [18].

References

1. Hunt, S., Abraham, W., Chin, M., Feldman, A., Francis, G., Ganiats, T., Jes-sup, M., Konstam, M., Mancini, D., Michl, K., Oates, J., Rahko, P., Silver, M., Stevenson, L., Yancy, C.: ACC/AHA2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure). (American College of Cardiology Web Site. Available at: <http://www.acc.org/clinical/guidelines/failure//index.pdf>)

2. Wooldridge, M.J., Jennings, N.R.: Intelligent agents: Theory and practice. *Knowledge Engineering Review* **10**(2) (1995) 115–152
3. Dominguez, D., Grasso, F., Miller, T., Serafin, R.: PIPS: An Integrated Environment for Health Care Delivery and Healthy Lifestyle Support. In: 4th Workshop on Agent applied in Healthcare, ECAI2006, Riva Del Garda (2006)
4. Goczyla, K., Grabowska, T., Waloszek, W., Zawadzki, M.: The knowledge cartography - a new approach to reasoning over description logics ontologies. [19] 293–302
5. Zambonelli, F., Jennings, N.R., Wooldridge, M.: Developing Multiagent Systems: The Gaia Methodology. *ACM Transactions on Software Engineering Methodology* **12**(3) (2003) 317–370
6. Booth, D., Haas, H., McCabe, F.: Web Services Architecture. <http://www.w3.org/TR/ws-arch/>(2005)
7. Battle, S., Bernstein, A., Boley, H., Grosz, B., Gruninger, M., Hull, R., Kifer, M., Martin, D., McIlraith, S., McGuinness, D., Su, J., Tabet, S.: Semantic Web Services Language (SWSL). <http://www.w3.org/Submission/SWSF-SWSL/> (2005)
8. OWL-S: OWL Web Service Ontology. (<http://www.daml.org/services/owl-s/>)
9. WSML: Web Service Modeling Language. (<http://www.wsml.org/>)
10. WSMO: Web Service Modeling Ontology. (<http://www.wsmo.org/>)
11. Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., Fensel, D.: Web Service Modeling Ontology. *Applied Ontology* **1**(2) (2005) 77 – 106
12. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible markup language (xml) 1.0 (third edition). (W3C Recommendation, <http://www.w3.org/TR/2004/REC-xml-20040204/>)
13. BPEL4WS: Business process execution language for web services. (<http://www-128.ibm.com/developerworks/library/specification/ws-bpel/>)
14. Larson, J.A., Raman, T., Raggett, D.: W3C Multimodal Interaction Framework. <http://www.w3.org/TR/mmi-framework/>, <http://www.w3.org/TR/mmi-arch/> (2003)
15. Barnett, J., Bodell, M., Raggett, D., Wahbe, A.: Multimodal Architecture and Interfaces. <http://www.w3.org/TR/mmi-arch/> (2006)
16. WURFL: Wireless Universal Resource File. (<http://wurfl.sourceforge.net/index.php>)
17. VoiceXML: Voice eXtensible Markup Language. (<http://www.voicexml.org/>)
18. Berners-Lee, T.: Xml and next generation web publishing. <http://www.w3.org/Talks/1998/0318-Seybold-timbl/overview.htm> (1998)
19. Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Stuller, J., eds.: SOFSEM 2006: Theory and Practice of Computer Science, 32nd Conference on Current Trends in Theory and Practice of Computer Science, Merín, Czech Republic, January 21-27, 2006, Proceedings. In Wiedermann, J., Tel, G., Pokorný, J., Bieliková, M., Stuller, J., eds.: SOFSEM. Volume 3831 of Lecture Notes in Computer Science., Springer (2006)

Author Index

- Abarca, Maria G. II-1976
Abdellatif, Takoua I-30
Abelló, A. I-40
Abels, Sven I-813
Agha, Gul I-32
Aguilera, Felipe II-1305
Ahlemann, Frederik I-813
Aларcon, Rosa A. II-1305, II-1976
Albani, Antonia I-763
Albertoni, Riccardo II-1863
Allen, Gabrielle I-52
Alvarez, Francisco I-658
Amara, Nejla II-1934
An, Dong Un I-50
Andjomshoaa, Amin II-1796
Anjewierden, Anjo I-279
Ariza, César II-1884
Armac, Ibrahim II-1315
Autiero, Ciro I-975
Avanes, Artin I-10
Avola, Danilo I-904
Azizah, Fazat Nur II-1221
- Backhaus, Michael I-669
Bagüés, Susana Alcalde II-1347
Baik, Doo-Kwon II-1844
Bakema, Guido II-1221
Bakillah, Mohamed II-1658
Balley, Sandrine II-1703
Balsters, Herman II-1201
Barria, Rodrigo II-1976
Barrio, Rubén I-342
Bartlang, Udo I-28
Bédard, Yvan II-1658
Behrens, Nils I-874
Bellatreche, Ladjel I-48
Ben Ahmed, Mohamed II-1028
Berlanga, Rafael II-1062
Bernard, Thibault II-1530
Bi, Jun II-1500
Bimonte, Sandro II-1596
Birney, Ewan I-710
Bittner, Thomas II-1626
Bodenstaff, Lianne I-802
- Boisson, Paul II-1712
Bolchini, Cristiana II-1986
Bollen, Peter II-1231
Bornhövd, Christof I-10
Borrell, Joan I-415
Böse, Joos-Hendrik II-1540
Bosque, José Luis I-78
Bothorel, Cécile I-290
Boyd, Colin I-616
Braun, Peter I-985
Brauner, Daniela F. I-46
Breit, Timo M. I-679
Bröcker, Lars I-44
Brouard, Thierry I-489
Buchanan, George II-1944
Bucher, Bénédicte II-1703
Bui, Alain II-1530
Buvarp, Tor Erik I-479
- Cai, Linlin II-1566
Calvi, Licia I-208, II-1956
Camossi, Elena II-1863
Campbell, Roy II-1337
Cao, Yang II-1480
Cao, Yiwei I-956
Cardot, Hubert I-489
Carminati, Barbara II-1734
Carpenter, M. I-15
Carver, Andy II-1201
Casanova, Marco A. I-46
Caschera, Maria Chiara I-904
Castellà-Roca, Jordi I-415
Castillo, Sergio I-415
Cauchie, Stéphane I-489
Celentano, Augusto I-916
Celestino Jr., Joaquim I-605
Ceravolo, Paolo II-1044, II-1825
Cerri, Stefano A. I-136
Çetintemel, Uğur II-1380
Challiol, Cecilia II-1914
Chandershekarapuram,
Appadodharana II-1367
Chang, Elizabeth II-1724, II-1765,
II-1806, II-1814, II-1835

- Chang, Po-Hao I-32
 Chatti, Mohamed Amine I-956
 Chatzigeorgiou, Alexander I-554
 Chen, Yi I-730
 Cheng, Jingde II-1264
 Cheong, Taesu II-1325
 Cho, Chulho I-720
 Choi, Eun Young I-362
 Choi, Kyu Young I-537
 Choi, Won-Hyuck II-1854
 Christiaens, Stijn I-199, I-689,
 II-1100, II-1191
 Chung, Seung Jong I-50
 Citro, Sandy I-21
 Clerc, Stéphane II-1712
 Clerckx, Tim I-884
 Coenen, Tanguy I-189
 Coessens, Bert I-689
 Cohn, Anthony G. II-1636
 Collazos, Cesar A. II-1305
 Coninx, Karin I-884, II-1966
 Corcho, Oscar I-36
 Corin, Ricardo I-626
 Coronato, Antonio II-1274
 Coulet, Adrien I-648
 Coulson, Geoff II-1490
 Covington, Michael J. II-1996
 Craig, Barbara I-237
 Crampes, Michel II-1050
 Croft, David I-710
 Curland, Matthew II-1181

 da Silva, Alberto Rodrigues I-824
 da Silva, Henrique J.A. I-404
 Dai, Tran Thanh I-585
 Damiani, Ernesto II-1044, II-1825
 Danger, R. II-1062
 Dawyndt, Peter I-740
 De Baer, Peter I-126
 De Baets, Bernard I-740
 De Bo, Jan II-1100
 de Bono, Bernard I-710
 de Buenaga, Manuel I-658
 de Castro, Valeria I-107
 De Cindio, Fiorella I-227
 De Decker, Bart I-636
 De Leenheer, Pieter II-1191
 De Maeyer, Philippe II-1636
 De Martino, Monica II-1863
 de Moor, Aldo I-247

 De Moor, Bart I-689
 De Pietro, Giuseppe II-1274
 De Vos, Paul I-740
 Dehainsala, Hondjack I-48
 dela Cruz, Necito II-1160
 Delgado Kloos, Carlos I-392
 Delgado, Jaime I-312, I-342
 Desconnets, Jean-Christophe II-1712
 Deters, Ralph I-19
 D'Eustachio, Peter I-710
 Devignes, Marie-Dominique I-648
 di Carlo, Vladimiro Scotto I-975
 Diallo, Gayo I-699
 Dillon, Tharam S. II-1724, II-1765,
 II-1806 II-1814, II-1835
 Donnelly, Maureen II-1626
 Dorn, Christoph II-1904
 dos Santos, André I-605
 Duan, Zhenhua II-1480
 Dubus, Jérémy I-26
 Dugenie, Pascal I-136
 Dustdar, Schahram II-1904
 Duvert, Frederic I-136
 Dyachuk, Dmytro I-19

 Eberhardt, Andreas I-763
 Eder, Johann I-5, I-42
 Efimova, Lilia I-279
 Eichner, Hannes I-5
 El Mahdi, Asmae I-301
 Eloff, J.H.P. I-322
 Endo, Takumi II-1264
 Esposito, Massimo II-1274
 Estevez-Tapiador, Juan M. I-352
 Evripidou, Paraskevas II-1285

 Feki, Ines II-1439
 Feldhofer, Martin I-372
 Fernandez, Marcial I-605
 Ferrari, Elena II-1734
 Ferreira, Diogo R. I-844
 Ferrer Gomila, Josep L. I-460
 Ferri, Fernando II-1670
 Fiumara, Giacomo I-227
 Flauzac, Olivier II-1530
 Folliot, Bertil II-1254
 Foo, Ernest I-616
 Formica, Anna II-1670
 Fortier, Andrés II-1914
 Foth, Marcus I-171

- Fournigault, Mike I-527
 Frampton, Keith I-34
 Frattini, Giovanni I-975
 Freire, Mário M. I-404
 Fresta, Giuseppe I-936
 Froihofer, Lorenz II-1510
 Fu, Yingfang I-575
 Fugazza, Cristiano II-1044
 Fuller, David II-1976
 Fürst, Frédéric I-38
- Gafurov, Davrondzhon I-479
 Galindo, David I-626
 Gan, Tian II-1566
 Gangemi, Aldo II-1012
 García-Alfaro, Joaquín I-415
 García Martínez, Alberto I-392
 García, Roberto I-40, II-1745
 Garrido, José L. I-863
 Gausmann, Oliver I-763
 Gaved, Mark I-171
 Gendarini, Domenico I-181
 Geoffroy, Nicolas II-1254
 Gevers, Steven I-636
 Giannini, Franca II-1863
 Gil, Rosa I-40, II-1745
 Gillespie, Marc I-710
 Gladt, Matthias II-1510
 Goeschka, Karl M. II-1510
 Goldfarb, Ilia I-926
 Gómez-Pérez, Asunción I-36
 González-Tablas, Ana Isabel II-1755
 Gopinath, Gopal I-710
 Gordillo, Silvia II-1914
 Gorton, Ian I-23
 Goto, Yuichi II-1264
 Gramoli, Vincent II-1470
 Grefen, Paul I-834
 Grifoni, Patrizia I-904, II-1670
 Grześkowiak, Maciej I-439
 Gu, Ming I-469
 Guerrero, Luis A. II-1305
 Gurstein, Michael I-301
- Hacid, Mohand-Said I-12
 Hahn, Axel I-813
 Hajduczenia, Marek I-404
 Halkidis, Spyros T. I-554
 Halpin, Terry II-1181, II-1201
 Han, Saeyoung I-55
- Han, Zongfen II-1776
 Hansen, Joe II-1160
 Hansen, Torben II-1402
 Hauck, Franz J. I-28
 Hausmann, Kevin I-813
 He, Jingsha I-575
 Hendrix, Marcel I-834
 Hernández Ernst, Vera II-1606
 Hernandez-Castro, Julio Cesar I-352
 Herre, Heinrich I-669
 Herrero, Pilar I-68, I-78, I-946
 Hieu, Cao Trong I-585
 Hinze, Annika II-1944
 Hirmer, Stephan I-52
 Hoang, Hanh Huu II-1796
 Hoareau, Didier I-30
 Hoehndorf, Robert I-669
 Hoepman, Jaap-Henk I-626
 Hong, Choong Seon I-585
 Hoppenbrouwers, S.J.B.A. (Stijn)
 II-1128, II-1138, II-1242
 Houben, Geert I-884
 Huang, Zewu II-1776
 Huguet Rotger, Llorenç I-460
 Hurtado, M. Visitación I-863
 Hussain, Farookh Khadeer II-1724,
 II-1765
 Hussain, Omar Khadeer II-1765
 Hutanu, Andrei I-52
 Hwang, Iksoon I-595
 Hwang, Soonhee II-1038
- Inácio, Pedro R.M. I-404
- Janowicz, Krzysztof II-1681
 Jans, Greet II-1956
 Jassal, Bijay I-710
 Jayawickrama, Wipul I-565
 Jeong, Dongwon II-1844
 Jeong, Jongil I-720, II-1357
 Jelic, Gordan I-97, I-752
 Jimenez-Ruiz, Ernesto II-1062
 Jin, Hai II-1776
 Jin, Hoon I-90
 Jing, Yixin II-1844
 Jlaiel, Nahla II-1028
 Jo, Heasuk I-331
 Jonquet, Clement I-136
 Jung, Youngim II-1038

- Kaiser, Hartmut I-52
 Kangasharju, Jussi I-894
 Kapitza, Rüdiger I-28
 Keet, C. Maria II-1118
 Kelso, Janet I-669
 Kenis, Dirk I-189
 Kenn, Holger I-874
 Kermarrec, Anne-Marie II-1470
 Kern, Steffen I-985
 Kerremans, Koen I-126
 Kettani, Driss I-301
 Kiani, Ali I-8
 Kim, Beob Kyun I-50
 Kim, Beomjoon I-595
 Kim, Bo Man I-537
 Kim, Chang Han I-382
 Kim, Chang-Soo I-792
 Kim, In-Cheol I-90
 Kim, Jae-Chul I-966
 Kim, Jinhung II-1844
 Kim, Jong-Woo I-792
 Kim, Ju-Wan I-966
 Kim, Ju-Yeon I-792
 Kim, Kyoung Hyun I-362
 Kim, Marie II-1325
 Kim, Seungjoo I-331
 Kirlidog, Melih I-257
 Kistijantoro, A.I. II-1555
 Klamma, Ralf I-956
 Ko, Eun Jung II-1948
 Kodratoff, Yves II-1107
 Kondratova, Irina I-926
 Konstantas, Dimitri II-1924
 Kunzmann, Christine II-1078
 Kwasnikowska, Natalia I-730
 Kwon, Hyuk-Chul II-1038

 Lacroix, Zoé I-730
 Lagerspetz, Eemil II-1894
 Lam, Herman I-780
 Lam, Kwok-Yan I-469
 Land, Martin Op't II-1419
 Lanubile, Filippo I-181
 Lau, Sian Lun II-1894
 Laurini, Robert II-1693
 Le, Khanh Vinh I-23
 Lee, Dong Hoon I-362, I-537
 Lee, Hyung Jik II-1948
 Lee, Jae-Jo I-585
 Lee, Jai-Ho I-966

 Lee, Jeun Woo II-1948
 Lee, Minsoo I-780
 Lee, Su Mi I-362
 Lee, Wookey II-1873
 Lee, Youngsook I-508
 Lee, Yunho I-331
 Leida, Marcello II-1825
 Leng, Xiaoxiang II-1500
 Leone, Giuseppe I-975
 Lever, Ryan II-1944
 Lewis, Suzanna I-710
 Li, Guorui I-575
 Liardet, Pierre-Yvan I-527
 Libourel, Thérèse II-1703, II-1712
 Lim, Jongin I-382
 Lim, Seungkil II-1873
 Lindeman, L. (Leonie) II-1242
 Liu, Chengfei I-1
 Liu, Tong I-247
 Liu, Y. II-1391
 Liu, Yan I-23
 Llorente, Silvia I-312
 Locatelli, Paolo II-1088
 Loebe, Frank I-669
 López-Cima, Angel I-36
 López, Guillermo I-107
 Lopez, Karla II-1693
 Luyten, Kris I-884, II-1966
 Lv, Ertao II-1480

 Madureira, Ricardo I-269
 Mahéo, Yves I-30
 Maitra, Anutosh II-1586
 Malinowski, Elzbieta II-1616
 Manset, David II-1062
 Marcante, Andrea I-936
 Marchi, Massimo I-227
 Marcos, Esperanza I-107
 Marino, Daniela I-995
 Marinoni, Clementina II-1088
 Marquis-Ogez, Emilie I-290
 Marshall, M. Scott I-679
 Martin, Miquel II-1894
 Martins, Paula Ventura I-824
 Matias, Ignacio R. II-1347
 Matthews, Lisa I-710
 Matthys, Eiblin I-189
 Mawlood-Yunis, Abdul-Rahman
 II-1021
 McClatchey, Richard II-1062

- McGovern, Jim I-21
 McIver Jr., William I-149
 Meersman, Robert I-689, II-1191
 Mehandjiev, N.D. I-15
 Melgara, Marcello I-995
 Mendez, Gonzalo I-946
 Merle, Philippe I-26
 Merzky, Andre I-52
 Milidiú, Ruy L. I-46
 Millerat, Jean II-1894
 Milleret-Raffort, Françoise II-1693
 Minout, Mohammed II-1648
 Miquel, Maryvonne II-1596
 Mitre, Hugo A. II-1755
 Moerman, Ingrid II-1966
 Monteiro, Paulo P. I-404
 Monti, Marina II-1863
 Moreau, Yves I-689
 Morgan, G. II-1555
 Morgan, Tony II-1201, II-1211
 Morkel, T. I-322
 Mosler, Christof II-1315
 Mossop, Dan I-517
 Mostafavi, Mir Abolfazl II-1658
 Mostefaoui, Achour II-1470
 Moulin, Bernard I-301
 Mühlhäuser, Max I-894
 Muñoz Merino, Pedro J. I-392
 Muñoz Organero, Mario I-392
 Mussio, Piero I-936
- Naisbitt, Jeffrey II-1337
 Nam, Junghyun I-508
 Napoli, Amedeo I-648
 Naudts, Dries II-1966
 Navarro, Guillermo I-415
 Neutens, Tijs II-1636
 Neyem, Andres II-1305
 Nicolai, Tom I-874
 Nieto, Juan Manuel González I-616
 Noguera, Manuel I-863
 Norbistrath, Ulrich II-1315
 Norta, Alex I-834
 Norton, Barry I-58
 Nurmi, Petteri II-1894
- O'Grady, M.J. II-1391
 O'Hare, G.M.P. II-1391
 Oliva, M. I-40
 Oliveira, Rui II-1520
- Oliveri, Elisabetta I-936
 Olivier, M.S. I-322
 Osrael, Johannes II-1510
- Paal, Stefan I-44
 Padula, Marco I-936
 Paletta, Mauricio I-68
 Panayiotou, Christoforos II-1295
 Papadopoulos, George A. I-17
 Paraire, Jordi II-1062
 Park, Jaesung I-595
 Park, Jong-Hyun I-966
 Park, Sungyong I-55
 Park, Young-Ho I-382
 Pascoe, Jason II-1884
 Paspallis, Nearchos I-17
 Payeras-Capellà, M. Magdalena I-460
 Pedrinaci, Carlos I-58
 Peng, Zhuo II-1480
 Pepels, Betsy II-1170
 Perdrix, F. I-40
 Perego, Andrea II-1734
 Pereira, José II-1520
 Perepletchikov, Mikhail I-34
 Pérez, María S. I-78
 Peris-Lopez, Pedro I-352
 Pernici, Barbara II-1088
 Perramon, Xavier I-342
 Petric, Ana I-752
 Petriccione, Pierpaolo I-975
 Phan, Raphael Chung Wei I-425
 Pichler, Horst I-5
 Pierra, Guy I-48
 Piprani, Baba II-1148
 Pittarello, Fabio I-916
 Plasmeijer, Rinus II-1170
 Podobnik, Vedran I-97, I-752
 Poortinga, Remco II-1894
 Porter, Barry II-1490
 Pose, Ronald I-517
 Post, Lennart I-679
 Proper, H.A. (Erik) II-1128, II-1138,
 II-1170, II-1242
 Provetti, Alessandro I-227
 Prüfer, Kay I-669
- Qi, Jian-Jun II-1480
 Qi, Zhichang I-116
 Quintarelli, Elisa II-1986
 Quix, Christoph II-1566

- Rabat, Cyril II-1530
 Radevski, Vladimir II-1068
 Rafanelli, Maurizio II-1670
 Ragia, LEMONIA II-1566
 Rajugan, R. II-1814
 Ramos, Benjamín II-1755
 Ranwez, Sylvie II-1050
 Ranwez, Vincent II-1050
 Rashbass, Jem I-3, II-1551
 Raynal, Michel II-1470
 Razali, Ermaliza I-425
 Rechberger, Christian I-372
 Reekie, Colette I-546
 Reichert, Manfred I-802
 Ribagorda, Arturo I-352, II-1755
 Ries, Sebastian I-894
 Rios, Alfonso II-1062
 Ripamonti, Laura I-227
 Robert-Inacio, Frédérique I-527
 Rocca, Fabio I-995
 Roche, Mathieu II-1107
 Rodrigues, Helena II-1884
 Rodrigues, Luís II-1520
 Roitman, Haggai II-1429
 Romano, Luigi I-975, I-995
 Roos, Marco I-679
 Rossak, Wilhelm I-985
 Rossi, Gustavo II-1914
 Ryan, Caspar I-21, I-34
 Ryu, JeHyok II-1380
- Sáenz, Fernando I-658
 Salden, Alfons II-1894
 Salgado, Ana Carolina II-1576
 Salvadores, Manuel I-78
 Samaras, George II-1295
 Sanna, Alberto I-995
 Santoro, Nicola II-1021
 Sanz, Ismael II-1062
 Sastry, Manoj R. II-1996
 Schall, Daniel II-1904
 Schmidt, Andreas II-1078
 Schmidt, Esther I-710
 Schmidt, Holger I-28
 Selim, Mohammad Reza II-1264
 Serafin, Riccardo I-995
 Sericola, Bruno II-1470
 Shand, Brian I-3, II-1551
 Shin, Dongil I-720, II-1357
 Shin, Dongkyoo I-720, II-1357
- Shiri, Nematollaah I-8
 Shrivastava, S.K. II-1555
 Sidhu, Amandeep S. II-1835
 Silva, Manuel I-269
 Simões, Dora I-269
 Simonet, Ana I-699
 Simonet, Michel I-699
 Simonovich, Daniel I-854, II-1409
 Smail-Tabbone Malika I-648
 Snekkenes, Einar I-479
 Soares, António Lucas I-269
 Sohn, Jaeeui I-55
 Song, Young-Ho II-1854
 Sonnante, Leonardo I-227
 Sorathia, Vikram II-1586
 Souza, Damires II-1576
 Srirama, Satish Narayana I-956
 Stalker, I.D. I-15
 Stein, Lincoln I-710
 Stephanides, George I-554
 Stillman, Larry I-237
 Strasunskas, Darijus II-1786
 Strickmann, Jan I-813
 Su, Stanley Y.W. I-780
 Suárez-Figueroa, María Carmen I-36
 Sun, Hong-Wei I-469
 Sun, Jia-Guang I-469
 Suomela, Jukka II-1894
 Supino, Gianluca I-975
 Sutterer, Michael II-1894
- Taïani, François II-1490
 Tamani, Electra II-1285
 Tan, Chik How I-450
 Tanasescu, Adrian I-12
 Tchounikine, Anne II-1596
 Tedesco, Patricia II-1576
 Teglia, Yannick I-527
 Temmerman, Rita I-126
 Terlouw, Linda II-1450
 Thaler, Andreas II-1540
 Thomas, Gaël II-1254
 Tibben, William I-160
 Tjoa, A. Min II-1796
 Tomassen, Stein L. II-1460, II-1786
 Torres, Víctor I-312
 Trémeau, Alain I-527
 Trichet, Francky I-38, II-1068
 Tritilanunt, Suratose I-616

- Trog, Damien II-1191
 Trzec, Krunoslav I-97
- Valdivielso, Carlos Fernandez II-1347
 Valladares, Ramon I-946
 van Bommel, P. II-1128, II-1138
 Van Brabant, Bart I-740
 Van Damme, Céline I-189
 Van de Weghe, Nico II-1636
 van der Weide, Th.P. II-1128, II-1138
 van Halteren, Aart II-1924
 Vaquero, Antonio I-658
 Vassilaras, Spyridon II-1367
 Vastrik, Imre I-710
 Vereecken, Jan II-1191
 Verlinden, Ruben I-689, II-1100
 Villerd, Jean II-1050
 Visagie, Johann I-669
 Viviani, Marco II-1825
 Vogiatzis, Dimitrios II-1367
 von Solms, Basie I-546
 Vrandečić, Denny II-1012
- Wac, Katarzyna II-1924
 Wehrle, Pascal II-1596
 Weigand, Hans I-218
 Weiler, Andrew II-1337
 Weiss, Michael II-1021
- Whitworth, Brian I-247
 Wiggisser, Karl I-42
 Winters, Frederik II-1966
 Witlox, Frank II-1636
 Wombacher, Andreas I-802
 Won, Dongho I-331, I-508
 Wongthongtham, P. II-1806
 Wu, Guanming I-710
- Ximenes, Pablo I-605
- Yang, Yun I-1
 Yoneki, Eiko I-874
 Yoo, Kee-Young I-499
 Yoon, Aesun II-1038
 Yoon, Eun-Jun I-499
 Youn, Taek-Young I-382
 Yovanof, Gregory S. II-1367
 Yuan, Pingpeng II-1776
- Zargayouna, Haïfa II-1934
 Zeidler, Andreas II-1347
 Zhang, Miao II-1500
 Zhao, Xiaohui I-1
 Zhong, Duhang I-116
 Ziekow, Holger I-10
 Zimányi, Esteban II-1616, II-1648

Erratum

LNCS 4277 Editorial

In an earlier version by mistake the volume editors have been stated instead of the authors of each paper. The current version is the correct source for any reference made to a paper included in the OTM 2006 Proceedings LNCS 4275-4278.