

Witold Abramowicz  
Dalia Kriksciuniene  
Virgilijus Sakalauskas (Eds.)

LNBIP 117

# Business Information Systems

15th International Conference, BIS 2012  
Vilnius, Lithuania, May 2012  
Proceedings

 Springer

Lecture Notes  
in Business Information Processing

117

Series Editors

Wil van der Aalst

*Eindhoven Technical University, The Netherlands*

John Mylopoulos

*University of Trento, Italy*

Michael Rosemann

*Queensland University of Technology, Brisbane, Qld, Australia*

Michael J. Shaw

*University of Illinois, Urbana-Champaign, IL, USA*

Clemens Szyperski

*Microsoft Research, Redmond, WA, USA*

Witold Abramowicz  
Dalia Kriksciuniene  
Virgilijus Sakalauskas (Eds.)

# Business Information Systems

15th International Conference, BIS 2012  
Vilnius, Lithuania, May 21-23, 2012  
Proceedings

 Springer

Volume Editors

Witold Abramowicz  
Poznań University of Economics  
Poznań, Poland  
E-mail: w.abramowicz@kie.ue.poznan.pl

Dalia Kriksciuniene  
Vilnius University  
Kaunas, Lithuania  
E-mail: dalia.kriksciuniene@khf.vu.lt

Virgilijus Sakalauskas  
Vilnius University  
Kaunas, Lithuania  
E-mail: virgilijus.sakalauskas@khf.vu.lt

ISSN 1865-1348  
ISBN 978-3-642-30358-6  
DOI 10.1007/978-3-642-30359-3  
Springer Heidelberg Dordrecht London New York

e-ISSN 1865-1356  
e-ISBN 978-3-642-30359-3

Library of Congress Control Number: 2012937368

ACM Computing Classification (1998): J.1, H.4, H.3

© Springer-Verlag Berlin Heidelberg 2012

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*Typesetting:* Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))



# Preface

The International Conference on Business Information Systems (BIS 2012) held during May 21–23, 2012, in Vilnius, Lithuania, was the 15th in a series of BIS conferences. The series has been widely recognized by professionals as a forum for the exchange and dissemination of topical research in the development, implementation, application and improvement of computer systems for business processes and information management.

As usual, the conference papers covered a broad range of BIS topics. This year again the topics related to business processes were the most popular. The proceedings start with an invited methodological paper by Prof. Buhl about the strengths and weaknesses of the information systems (IS) and business and information systems engineering (BISE) communities. It is then followed by the two sessions on business process discovery and business process verification, touching on important issues of process mining, monitoring and consistency checking. The session on service architectures is also one of the canonical sessions during BIS. The process-related part is concluded with the session on collaborative BIS. In all, six papers were devoted to classic topics such as data management and Web search. Several authors researched applications of BIS in finance. Finally, there were several niche papers that were devoted to specific BIS issues.

Altogether, one invited talk and a set of 26 best evaluated papers illustrating these trends were selected out of 70 submissions for presentation during the main event, grouped into nine sessions. The Program Committee consisted of almost 100 members, who carefully evaluated all the submitted papers. Each submission was reviewed on average by 3.3 Program Committee members.

The conference was honored by three keynote speakers: Ajith Abraham (Machine Intelligence Research Labs, USA), Emilio Corchado (University of Burgos, Spain), and Gerald Quirchmayr (University of Vienna, Austria).

May 2012

Witold Abramowicz  
Dalia Kriksciuniene  
Virgilijus Sakalauskas

# Organization

## BIS 2012 was organized by

- Vilnius University, Kaunas Faculty of Humanities, Department of Informatics, and
- Poznań University of Economics, Department of Information Systems

## Program Chairs

Witold Abramowicz  
Dalia Kriksciuniene  
Virgilijus Sakalauskas

## Program Committee

Witold Abramowicz	Poznań University of Economics, Poland
Dimitris Apostolou	University of Piraeus, Greece
Morad Benyoucef	University of Ottawa, Canada
Hans U. Buhl	Augsburg University, Germany
Michelangelo Ceci	University of Bari, Italy
Wojciech Cellary	Poznań University of Economics, Poland
Jeng-Chung Chen	National Cheng Kung University, Taiwan
Dickson K.W. Chiu	Dickson Computer Systems, Hong Kong
Oscar Corcho	Universidad Politecnica de Madrid, Spain
Zhihong Deng	Peking University, China
Tommaso Di Noia	Technical University of Bari, Italy
Schahram Dustdar	Technical University of Vienna, Austria
Gintautas Dzemyda	Vilnius University, Lithuania
Suzanne Embury	University of Manchester, UK
Agata Filipowska	Poznań University of Economics, Poland
Bogdan Franczyk	University of Leipzig, Germany
Flavius Frasincar	Erasmus School of Economics, The Netherlands
Johann-Christoph Freytag	Humboldt University, Germany
Francesco Guerra	University of Modena, Italy
Jon Atle Gulla	Norwegian University of Science and Technology, Norway
Hele-Mai Haav	Tallinn University of Technology, Estonia
Axel Hahn	University of Oldenburg, Germany
Stephan Haller	SAP Research Zürich, Switzerland
Willem-Jan van den Heuvel	Tilburg School of Economics and Management, The Netherlands
Knut Hinkelmann	University of Applied Sciences Northwestern Switzerland, Switzerland

Marta Indulska	The University of Queensland, Australia
Marijn Janssen	Delft University of Technology, The Netherlands
Adam Jatowt	Kyoto University, Japan
Monika Kaczmarek	Poznań University of Economics, Poland
Tomasz Kaczmarek	Poznań University of Economics, Poland
Paweł J. Kalczyński	CSU Fullerton, USA
Yuriy Kharin	Belarusian State University, Belarus
Gary Klein	University of Colorado at Colorado Springs, USA
Ralf Klischewski	German University in Cairo, Egypt
Andrzej Kobylński	Warsaw School of Economics, Poland
Marek Kowalkiewicz	SAP Research Brisbane, Australia
Helmut Krcmar	TU München, Germany
Dalia Kriksciuniene	Vilnius University, Lithuania
Nor Laila Md Noor	Universiti Teknologi MARA, Malaysia
Daniel Lemire	Université du Québec à Montréal, Canada
Alexander Löser	Technische Universität Berlin, Germany
Qiang Ma	Kyoto University, Japan
Maria Mach	University of Economics in Katowice, Poland
Leszek Maciaszek	Macquarie University, Australia
Yannis Manolopoulos	Aristotle University, Greece
Florian Matthes	Technical University of Munich, Germany
Heinrich C. Mayr	University of Klagenfurt, Austria
Jan Mendling	Wirtschaftsuniversität Wien, Austria
Markus Nüttgens	University of Hamburg, Germany
Andreas Oberweis	University of Karlsruhe, Germany
Mitsunori Ogihara	University of Miami, USA
Sevgi Özkan	Middle East Technical University, Turkey
Marcin Paprzycki	Polish Academy of Sciences, Poland
Eric Paquet	National Research Council, Canada
Carlos Pedrinaci	The Open University, UK
Klaus Pohl	University of Duisburg-Essen, Germany
Jaroslav Pokorný	Charles University, Czech Republic
Elke Pulvermueller	University of Osnabrück, Germany
Manjeet Rege	Rochester Institute of Technology, USA
Ulrich Reimer	University of Konstanz, Germany
Gustavo Rossi	National University of La Plata, Argentina
Massimo Ruffolo	University of Calabria, Italy
Shazia Sadiq	The University of Queensland, Australia
Virgilijus Sakalauskas	Vilnius University, Lithuania
Sherif Sakr	University of New South Wales (UNSW), Australia
Demetrios Sampson	University of Piraeus, Greece
Jürgen Sauer	University of Oldenburg, Germany
Ulf Seigerroth	Jönköping University, Sweden

Gheorghe Cosmin Silaghi	Babes-Bolyai University of Cluj-Napoca, Romania
Elmar J. Sinz	University of Bamberg, Germany
Kilian Stoffel	University of Neuchâtel, Switzerland
Darijus Straszunas	Norwegian University of Science and Technology, Norway
Jerzy Surma	Warsaw School of Economics, Poland
Zdzisław Szyjewski	University of Szczecin, Poland
Sergio Tessaris	Free University of Bozen-Bolzano, Italy
Bernhard Thalheim	Universität Kiel, Germany
Barbara Thönssen	University of Applied Sciences Northwestern Switzerland, Switzerland
Robert Tolksdorf	Free University Berlin, Germany
Vassileios Tsetsos	University of Athens, Greece
Herna Viktor	University of Ottawa, Canada
Stefan Voß	Hamburg University, Germany
Mathias Weske	Hasso Plattner Institute for IT-Systems Engineering, Germany
Krzysztof Węcel	Poznań University of Economics, Poland
Anna Wingkvist	Linnaeus University, Sweden
Andreas Wombacher	University of Twente, The Netherlands
Yun Yang	Swinburne University of Technology, Australia
Qi Yu	Rochester Institute of Technology, USA
Slawomir Zadrozny	Polish Academy of Sciences, Poland
John Zeleznikow	Victoria University, Australia
Jozef Zurada	University of Louisville, USA

## Additional Reviewers

Abdullah, Syed Norris Hikmi	Neubert, Christian
Bewernik, Marc-Andre	Oro, Ermelinda
Dongus, Konrad	Roth, Sascha
Geuter, Juergen	Scheffler, Alexa
Hauder, Matheus	Schlachtbauer, Tobias
Jayawardene, Vimukthi	Schmieders, Eric
Krause, Felix	Setiawan, Mukhammad Andri
Manner, Julia	Sobczak, Andrzej
Metzger, Andreas	Ulfat-Bunyadi, Nelufar
Moldovan, Darie	

## Local Organization

Elżbieta Bukowska	Poznań University of Economics
Dalia Kriksciuniene (Co-chair)	Vilnius University
Szymon Łazaruk	Poznań University of Economics
Piotr Stolarski	Poznań University of Economics
Krzysztof Węcel (Co-chair)	Poznań University of Economics

# Table of Contents

## Invited Paper

Information Systems and Business & Information Systems Engineering: Status Quo and Outlook . . . . .	1
<i>Hans Ulrich Buhl and Martin Lehnert</i>	

## Business Process Discovery

Mining Constraints for Artful Processes . . . . .	11
<i>Claudio Di Ciccio and Massimo Mecella</i>	

Creating Sound and Reversible Configurable Process Models Using CoSeNets . . . . .	24
<i>Dennis M.M. Schunselaar, Eric Verbeek, Wil M.P. van der Aalst, and Hajo A. Raijers</i>	

Semantics-Based Business Process Model Similarity . . . . .	36
<i>Bernhard G. Humm and Janina Fengel</i>	

## Business Process Verification

Data- and Resource-Aware Conformance Checking of Business Processes . . . . .	48
<i>Massimiliano de Leoni, Wil M.P. van der Aalst, and Boudewijn F. van Dongen</i>	

An Approach for Consistent Delegation in Process-Aware Information Systems . . . . .	60
<i>Sigrid Schefer-Wenzl, Mark Strembeck, and Anne Baumgrass</i>	

Goal-Oriented Model-Driven Business Process Monitoring Using ProGoalML . . . . .	72
<i>Falko Koetter and Monika Kochanowski</i>	

## Service Architectures

Automatic Derivation of Service Candidates from Business Process Model Repositories . . . . .	84
<i>Henrik Leopold and Jan Mendling</i>	

Adaptability of Service Based Workflow Models: The “Chained Execution” Architecture ..... 96  
*Saida Boukhedouma, Zaia Alimazighi, Mourad Oussalah, and Dalila Tamzalit*

A Conceptual Model for Assessing the Benefits of Software as a Service from Different Perspectives ..... 108  
*Anisah Herdiyanti Prabowo, Marijn Janssen, and Joseph Barjis*

**Collaborative BIS**

Collaboration Infrastructure for the Learning Organization ..... 120  
*Keith Harrison-Broninski and Janne J. Korhonen*

IdeaWall: Mixed Mode Distributed Collaboration in Enterprise Environments ..... 132  
*Marek Kowalkiewicz*

A Novel Approach to Increase Efficiency of OSS/BSS Workflow Planning and Design ..... 142  
*Tetiana Kot, Andrey Reverchuk, Larysa Globa, and Alexander Schill*

**Data Management**

Recognition and Pseudonymization of Personal Data in Paper-Based Health Records ..... 153  
*Stefan Fenz, Johannes Heurix, and Thomas Neubauer*

A Detailed Process Model for Large Scale Data Migration Projects ..... 165  
*Klaus Haller, Florian Matthes, and Christopher Schulz*

Towards Automated *Generic* Electronic Flight Log Book Transfer ..... 177  
*Carsten Kleiner and Arne Koschel*

**Web Search Applications**

Authoring Processing Chains for Stream-Based Internet Information Retrieval Systems ..... 189  
*Philipp Katz, Marius Feldmann, Torsten Lunze, Sebastian Sprenger, and Alexander Schill*

On the Precision of Search Engines: Results from a Controlled Experiment ..... 201  
*Hasan Girit, Robert Eberhard, Bernd Michelberger, and Bela Mutschler*

Towards Semantic Search for Improved Decision-Making ..... 213  
*Darijus Strasunskas and Stein L. Tomassen*

**BIS in Finance**

Synthetic History for Exchange Traded Funds . . . . .	224
<i>Aistis Raudys, Lukas Sirvydis, and Karol Lisovskij</i>	
Dynamic Adaptive Algorithm Selection: Profit Maximization for Online Trading . . . . .	236
<i>Iftikhar Ahmad, Javeria Iqbal, and Günter Schmidt</i>	
Credit Risk Modeling of USA Manufacturing Companies Using Linear SVM and Sliding Window Testing Approach . . . . .	249
<i>Paulius Danenas and Gintautas Garsva</i>	

**Decision Support**

A Context Framework for Process-Oriented Information Logistics . . . . .	260
<i>Bernd Michelberger, Bela Mutschler, and Manfred Reichert</i>	
Making Recommendations for Decision Processes Based on Aggregated Decision Data Models . . . . .	272
<i>Razvan Petrusel and Paula Ligia Stanciu</i>	
Investigation of Performed User Activities in Overall Context with IT Analytical Framework . . . . .	284
<i>František Babič, Jozef Wagner, and Ján Paralič</i>	

**Specific BIS Issues**

Lightweight Certificates – Towards a Practical Model for PKI . . . . .	296
<i>Lukasz Krzywiecki, Przemysław Kubiak, Mirosław Kutylowski, Michał Tabor, and Daniel Wachnik</i>	
Control of Assistive Tools Using Voice Interface and Fuzzy Methods . . . . .	308
<i>Vytautas Rudzionis, Rytis Maskeliunas, and Tomas Rasyimas</i>	
<b>Author Index</b> . . . . .	319

# Information Systems and Business & Information Systems Engineering: Status Quo and Outlook\*

Hans Ulrich Buhl and Martin Lehnert

Research Center Finance & Information Management,  
University of Augsburg, 86135 Augsburg, Germany  
{hans-ulrich.buhl, martin.lehnert}@wiwi.uni-augsburg.de

**Abstract.** Although both communities share a common object of research, the Business and Information Systems Engineering (BISE) community from the German-speaking countries and the Information Systems (IS) community centered in North America have developed quite differently. The BISE community features promote connections with industry, attractive topics to students and practical relevance of publications. But due to various reasons numerous BISE researchers struggle with publications in top-ranked journals. While this weakness obviously is a strength of the IS community, we observe that the IS community struggles with its industry connections and enrollment numbers. What the global IS/BISE community needs is a more intense discourse that increases mutual understanding, creates awareness for the need for complementation, and ensures that the opportunity for complementation is seized. This paper offers insights on how by complementation both communities could mitigate some of their weaknesses and the global IS/BISE community could increase its success as a whole.

**Keywords:** Information Systems, Business and Information Systems Engineering, BISE, Critical Reflection, Scenario Analysis.

## 1 The Need and Opportunity for Complementation

This is a story about complementation. Although it may seem contradictory, the foremost task of one advocating complementation is segmentation. It seems even more contradictory if one considers that in general segmentation is unable to capture a complex spectrum of shades of grey – particularly if phenomena such as scientific communities are concerned. The reason, however, is simple: Without segmentation, differences remain opaque and rationales for complementation cannot be justified.

For increased contour of argument and with admitted oversimplification, we focus on the community from the German-speaking countries, i.e., Germany, Austria, and Switzerland, and the community from North America. We choose these communities

---

\* This paper is both a partially shortened and in some parts an extended version of the paper [5] Buhl, H.U., Fridgen, G., Müller, G., Röglinger, M.: Business and Information Systems Engineering: A Complementary Approach to Information Systems - What We Can Learn from the Past and May Conclude from Present Reflection on the Future. Journal of the Association for Information Systems, (2012).



because they have developed rather independently for a long time and prototypically epitomize different characteristics [1, 2]. These characteristics include sources of funding, teaching, predominating research paradigm, doctoral and post-doctoral qualification, and interplay with other disciplines. In our opinion, focusing on the communities from the German-speaking countries and North America is not too reductionist. On the one hand, some researchers pointed to similarities between the community from the German-speaking countries and other communities from Europe and Australasia [1, 3, 4]. On the other hand, the characteristics predominating in North America have been adopted by the vast majority of communities worldwide. A geographic segmentation seems appropriate because we are interested in the communities in their entirety and because any other set of segmentation criteria will result in oversimplification as well. Indeed, neither community is perfectly homogeneous. Some researchers from each community may feel closer to the other community regarding their individual approach and environment. The comparison of two large communities with a long tradition of high quality research allows us to determine notes for guidance to countries which only recently started out to participate in the international science scene.

Throughout this paper, we refer to the community from the German-speaking countries as BISE community. This is because the journal *Business & Information Systems Engineering (BISE)/WIRTSCHAFTSINFORMATIK* has been the community's primary publication outlet during the last fifty years and thus is a mirror of its evolution. As the characteristics predominating in North America have been adopted to a much higher extent than the characteristics of the BISE community, we refer to the North American community as the information systems (IS) community. Whenever we address all researchers dealing with information systems as object of research, we use the notion global IS/BISE community.

Both the IS and the BISE communities more and more think outside of the box and consider that a broad discussion on the opportunity for a better cooperation between both communities is advantageous. Among other things this is evidenced by the recent publication of the discussion [5] on the history of the journal BISE deriving some recommendations for both the IS and BISE communities. Also, an article [6] on learnings for the strategic information systems community from the experience of the BISE community is going to be published, with a main focus on the relationship to the industry. Apart from this intensifying exchange between the IS and BISE communities, there are an increasing number of authors who look beyond their home communities' noses [some examples are 1, 2, 7, 8, 9].

If both communities are willing to learn from each other, it will be possible to reduce the weaknesses of each individual community and to achieve a win-win setting for the global IS/BISE community. For new players on the market, like some of the Eastern European countries, the goal should be to adopt the rigor from the IS community and the relevance from the BISE community.

This article starts out with a short reflection of the status quo of the BISE community and compares the IS and BISE communities, the academic careers, the relation to industry, and to students (section 2). Based on the insight that the BISE and the IS community have the opportunity to make use of their complementary strengths, we discuss what might happen depending on whether this opportunity is seized or not (section 3). The paper concludes with recommendations from a BISE perspective that may serve as cornerstones for the transformation towards complementation (section 4).

## 2 A Comparison of the BISE and IS Communities

At the beginning of this chapter we start out with a short status quo of the journal *Business & Information Systems Engineering (BISE)/WIRTSCHAFTSINFORMATIK*. After that, the chapter provides a comparison of the BISE community consisting of the German-speaking countries, and the IS community centered in North America that has a strong international dissemination.

Driven by the “Wirtschaftswunder” (economic miracle) and the increasing opportunity for industry to adopt electronic computers, the BISE journal’s progenitor Elektronische Datenverarbeitung (Electronic Data Processing) was founded in 1959 by Hans Konrad Schuff, the executive manager of the first European software house mbp. Already at this time, the editorial board included editors from academia and industry [10]. After an eventful history the *WIRTSCHAFTSINFORMATIK* (the journal was renamed in the meantime) went through some challenging times before its 50<sup>th</sup> anniversary due to decreasing submissions and a dwindling subscriber base caused by higher scientific standards that had been established which made papers cumbersome to read for practitioners. Therefore, on the occasion of its 50th anniversary, *WIRTSCHAFTSINFORMATIK* implemented a strategic realignment and is henceforth complemented by the English-language e-journal *Business & Information Systems Engineering (BISE)*. The editorial board was extended by experts from the IS community who form bridges between both communities. Departments were established and staffed with editor teams from both communities. Editors from industry were kept as well. In addition, there was a consensus that a single journal cannot simultaneously satisfy the needs of international researchers and German-speaking practitioners. Therefore, the *Wirtschaftsinformatik & Management (WUM)* journal was launched to maintain knowledge exchange with industry – analogous to *MIS Quarterly Executive*. *WUM* inherited the practitioner-oriented sections of the scientific journals, developed them further, and provides management summaries of research papers. The connection between industry and academia was further strengthened by the fact that subscribers have access to all online archives no matter which of these journals he or she obtains in print.

The strategic realignment of the journal *BISE/WIRTSCHAFTSINFORMATIK* with its integrative approach - combining some strengths of both the IS and the BISE communities - was quite successful: It was announced as the first AIS Affiliated Journal just prior to *ICIS 2010*. In 2011, *BISE*’s full text downloads mounted up to 300 % compared to 2009 and the impact factor tripled within three years.

After this brief review of the changes in the journal, a comparison between the IS and BISE community follows. As outlined above by using the recent history of the BISE journal as example, the BISE community has been closely linked with industry since its very beginnings. Many of the features that characterize the BISE community today have been determined or at least influenced by its industry connections. On the other hand the IS community has a long history of publishing their scientific research results in top-journals. Mainly adhering to the natural science paradigm, theories are used for explanation and prediction [11, 12]. The link with industry, however, is not especially strong, and the added value for decision makers in companies is often rather low.

**Associated Community:** BISE chairs have their origin mostly from engineering and computer science schools and the self-conception of the BISE chairs follows the tradition of engineering science. Even if BISE chairs are located at business schools, they often receive considerably better staffing and funding as if they were located at a computer science or engineering school. Therefore, they use a design-oriented way to solve business problems and have a high involvement with industry. Peter Mertens [13], one of the BISE community's founders, postulates that researchers prove themselves in a decathlon of objectives. Almost half of these objectives require boundary spanning between academia and industry (e. g., conducting applied research projects, supporting start-ups and spin-offs, placing students as interns, and raising funds from industry). This model is crisis-proof and leaves more freedom to individual researchers as legitimation may be drawn from various sources. There is a reason why BISE researchers have recently earned a reputation as "happy souls" [2].

Numerous BISE professors expand their chairs to "scientific think tanks" of more than 20–30 research assistants that do both fundamental and applied research. In some cases, several professors team up and found research centers. Such think tanks and research centers feature a staff that is diverse enough to conduct (applied) research projects with various foci. They manage to maintain and increase research and project management competences in an environment where most of the staff drops out after 3–5 years due to the end of their doctorates.

According to our perception, many IS researchers are excellent in investigating the transformational power of IT and its impacts on individuals or teams [14]. They disclose general insights and document them as justified theories [15]. Their research is mostly explanation-oriented with the goal of predicting how society or parts of society interact with Information Systems. However, solving the problems of the industry is mostly not the main goal.

**Academic Careers:** The vast majority of doctoral students in the BISE community intentionally seek management careers after finishing their doctorate. Doctoral work therefore emphasizes analytical and project management skills, while training in research methods and writing skills has been secondary for a long time.

In the IS community PhD programs are geared to scientific careers, which is why researchers dispose of profound theoretical knowledge and have a high command of research methods. They are also inducted into a strong publishing and reviewing culture. Accordingly, numerous scholarly IS journals are recognized as standard setters with respect to methodological rigor and scholarly writing.

**Relation to the Industry:** As mentioned before, BISE is strongly connected to industry. Throughout its evolution, the BISE community maintained its focus on solving business problems by means of useful artifacts. The consequences manifest themselves in the community's sources of funding. Most chairs additionally employ an arbitrary number of research assistants funded by grants or applied research projects with industry. Altogether, approximately 44 % of research assistants are funded by industry [1].

In contrast, the IS community has separated from the business problems and applied research projects. The research is dedicated to general insights and to

documenting them as justified theories. The result is a “disconnect between the worlds of business and academia” [16]. The funding of IS chairs has gone down from “90% industry funding to 95% government funding” [17] over the last 20 years.

**Students:** As for degree programs, companies usually get involved both financially and by means of additional courses (e. g., project seminars, guest lectures, jointly supervised bachelor or master theses) to get acquainted with future graduates at an early stage. Hence, courses deal with topics of practical relevance, include cases from applied research projects, and are enriched by the researchers’ practical experience. To sum up, BISE degree programs are highly attractive. In Germany, for instance, annual BISE enrollments doubled from 2000 to 2010 (Federal Statistical Office 2011). Almost all universities from the German-speaking countries offer dedicated BISE degree programs.

On the other hand, in the IS community enrollment numbers are falling and courses are deleted from MBA programs at many universities [18, 19, 20, 21]. The consequence of very theoretical research is also a very theoretical education of students. Therefore, graduates of IS programs are mostly less attractive for industry than comparable graduates of BISE programs.

**Current Challenges:** Opinions on the respective other community are inspired more by anecdotes than by facts. The resulting prejudices can sometimes be read between the lines and are known by hearsay. For instance, “BISE is consulting!”, “IS is irrelevance at the highest stage!”, “BISE does everything that gets funded by industry!”, “IS publishes everything where data is available!”, “BISE has never shown results for the money invested in research!”, or “IS is no more than behaviorist research!”

From these exaggerate statements the current problems of the two communities can be summed up pointedly: BISE follows research that borders perilously on consulting, is addicted to technological fads, has a sloppy reviewing and quality culture, and lacks a long-term research agenda. Others criticize the substandard output of publications in top-ranked journals. In recent years, industry connection is at risk because universities and funding organizations increasingly impose incentives as well as assessment and tenure criteria that are rather exclusively based on publications in top-ranked journals. Instead the IS community struggles with its identity, legitimation, and industry connection [16, 22, 23, 24, 25, 26, 27].

If everything continues as before, both communities run into major problems. IS uses methodologically rigorous research, but lacks practical relevance. BISE has relevant applied research close to consulting, but lacks rigor. Still, both will claim to be rigorous and relevant.

### 3 What May Lie Ahead: Possible Scenarios

**Dinosaurs Heading Blindly towards Extinction or BISE Becomes Like IS:** On the one hand, Lyytinen et al. [9] argue that replicating the U.S. system would lead to more publications in top-ranked journals for BISE researchers. On the other hand, IS scholars discuss that “if European researchers are tempted to move away from their

practice-informing activities in a quest for U.S.-style research publications, that does not bode well for the European model of [IS] research” [8]. So what could be the consequence for BISE’s currently established system?

The title “doctor in BISE” will lose its hard gained reputation in industry and become less attractive for young academics, due to a necessary cut of the relation to industry and education of project management skills. The cut is required to make sure that the researcher can compete with U.S.-style doctoral programs. U.S.-style doctoral programs focus on a pure research career. Many BISE researchers are not trained in typical IS methods and publishing cultures. They usually have no interfaces to psychology, philosophy, or social science. Thus, building up U.S.-style doctoral programs will almost never directly lead to publication success. Instead, these programs will compete with the BISE typical scientific think tanks for young academics.

Second, if top journal publications keeps becoming the predominant criterion for grants, those grants will be given to few specialized “mile deep/inch wide lonesome cowboys” who exclusively focus on research and supervise only few doctoral students. Ironically, lonesome cowboys love their lonesomeness and apply for grants rather for the kudos than for the money. This renunciation of industry-related research will lead into a drop of BISE’s enrollments and private funding. As a result, BISE’s scientific think tanks, which are currently successful regarding the societal and economic impact of their research, will be lose the financial and human capital that is the fundament of their success. In a vicious circle, BISE’s enrollments and private funding will drop. The system as it used to be will die – so the fear – with 99 % of frustrated losers and 1 % of neurotic winners.

This loss of global diversity will also accelerate IS’ expected “downward spiral because of [...] increasing narrow-mindedness” [17]. IS will continue to lose enrollments, to be stuck in its identity crisis, and slowly become an endangered species.

Although some IS scholars complain about perpetually “lamenting the state of information systems as a discipline” [25], those tendencies worry many BISE researchers. They wonder: “Why should we let that happen? Why should we take IS as a role model, despite its problems in identity, enrollments, and relevance? BISE is successful as it is except for publications in journals, which no practitioner ever reads anyway. Why should we adopt the IS identity crisis?”

**A Split of Communities or Irrelevant vs. Ir-rigorous Researchers:** Each community might split into (even more) *distinct* sub-communities. IS-style scholars in German-speaking countries prefer to collaborate with their North American counterparts and not with their local colleagues. BISE-style researchers in North America prefer to collaborate with computer scientists or engineers who consequently begin to take over those fields of research. BISE and IS will be regarded as no more than fringe groups of computer science and sociology.

IS will be rigorous research, but lack relevance. BISE will be relevant applied research close to consulting, but lack methodological rigor. Over time, business schools will prefer to tenure IS scholars. Computer science or engineering schools will tenure BISE scholars. The opportunities for academic offspring to become

acquainted with the respective other perspective will be rare. Even journals might decide on their affiliation. There will be silence between both communities.

There may even be a break-up of universities. On the one hand, there would be purely publicly funded research universities with high scientific impact, e. g. in theoretical physics, mathematics, social science, or business research. But neither the researchers themselves nor their students would achieve business impact; business schools would even be harmful for management practices [28]. On the other hand, applied universities with a high portion of private funding would have strong engineering schools with high relevance, but no chance for public grants following international criteria. If at all, they could get grants from the ministries responsible for economics or technology.

Surely, any individual scholar could be happy within one of these scenarios as each kind of research – be it IS or BISE, rigorous or relevant, behavioral or design-oriented – would be allowed. The question is: What would be the long-term impact on the IS/BISE community, and even more importantly, on business and society? Science and industry becoming more and more independent obviously bears the risk of losing touch. Theory would develop models far from reality and business would substitute gut feelings for methodologically well-founded decisions. Neither model would be appealing for young academics who seek rigor *and* relevance as well as jobs at the intersection between business and research. Except for very few talents, IS and BISE would cease to exist. Maybe the last dinosaurs lived a happy life, too.

Before getting to our favored scenario, we have a quick look at two (not really serious) scenarios with a low probability.

**IS Becomes BISE:** Even if recent articles like Lee [12] or Gill and Bhattacharjee [8] are weak signals that IS scholars are becoming more and more aware that Europe offers different approaches that could be fruitful for the relevance of their research, this is not really a serious scenario.

**BISE and IS Switch Roles:** Having said that, if the “Americanization” of the European research landscape continues, there is a small, but positive probability for the opposite scenario, that BISE and IS simply “switch roles”. Thus we would probably have the same discussion the other way around in some dozen years.

**Towards IS and BISE as Complements:** In our opinion, only one scenario yields strong contributions to theory, business, and society: IS and BISE complement each other and make use of their strengths to cope with respective weaknesses and threats. BISE researchers strive for “giant leaps” to boldly answer relevant research questions no man has asked before. In contrast, typical IS journals value “incremental articles [that] focus on a single question based on an assumption ground that has been established elsewhere” [9].

Both approaches are required, however, both must be complementary and not mutually exclusive. After a revolutionary discovery, it must be possible to perform further developments evolutionarily. When you compare the characteristics of the IS and BISE communities, it becomes evident that more practical orientation in IS research and more theoretical foundation in BISE research, would be a useful

complement. For this purpose, it is necessary to raise the relevance of the issues in the IS and get a better methodological basis in students education in BISE. A respectful cooperation is essential for a close collaboration between the two ways of our discipline. Only then is it possible to share knowledge, complement strengths, and compensate weaknesses and threats. The priority objective is an integrated approach of the global IS/BISE community.

Henning Kagermann, former CEO of SAP, who actually holds a professorship in theoretical physics, draws an interesting parallel to the research culture in his original discipline [29]: Physics integrates mathematical modeling, experiments and empirical tests. In design science, there is an integration of methodologies, too. But there is one crucial difference: Typical definitions of design science require design and evaluation to be done by the *same* researchers and to be published jointly in *each* design-oriented paper. Theoretical physicists are not good experimental physicists and vice versa. Nevertheless, both respect one another and there exist a lot of research groups in physics where theoretical and experimental physicists team up to complement each other.

If all representatives of the IS and BISE community respect one another and neither would seriously doubt the others' strengths and mission, they could share enough knowledge to communicate their problems at hand. Complementarity between the IS and BISE communities, or in other words division of labor with defined interfaces, is necessary for contributing to theory, business, and society. Wouldn't that be an interesting perspective for the entire IS/BISE community?

## 4 Recommendations

The goal should be an integrative approach. It must make use of the respective IS and BISE strengths and the weaknesses must be compensated. An exchange of both communities is essential. From this, other communities could learn as well and use the experience of such two large communities for their benefit. The crucial point is not to repeat mistakes, but instead to prevent already-known weaknesses of the IS and BISE scientific system and to adapt the strengths of the whole IS/BISE community.

This way, developing scientific communities in emerging countries will reduce the distance to the leading nations significantly faster and increase the probability to become a major player in the global science market.

From our perspective, both communities have to strive for a *common* vision over the next years and decades. Three things are certain: First, there has to be an even more intensive discourse within and between both communities. Second, multiple stakeholders will have to act or be forced to act. Third, deliberate adaptation will be necessary to avoid losing parts of the communities. Accomplishing this is not going to be a walk in the park, but will pay off in the long run!

Summarizing, comparing the BISE community's strengths and weaknesses due to its industry relations in funding, teaching, and research with the perceived strengths and weaknesses of the IS community, we conclude that the BISE community can offer experience in areas where the IS community seems to have problems and vice versa. Thus, both communities have the opportunity to complement each other.

## References

1. Frank, U., Schauer, C., Wigand, R.T.: Different Paths of Development of Two Information Systems Communities: A Comparative Study Based on Peer Interviews. *Communications of the Association for Information Systems* 22(1), 391–412 (2008)
2. Junglas, I., Niehaves, B., Spiekermann, S., Stahl, B., Weitzel, T., Winter, R., Baskerville, R.: The inflation of academic intellectual capital: the case for design science research in Europe. *European Journal of Information Systems* 20(1), 1–6 (2011)
3. Loos, P., König, W., Österle, H., De Marco, M., Pastor, J.A., Rowe, F.: National Research and International Competitiveness—An Antinomy? *Business & Information Systems Engineering* 2(4), 249–258 (2010)
4. Winter, R.: Design science research in Europe. *European Journal of Information Systems* 17(5), 470–475 (2008)
5. Buhl, H.U., Fridgen, G., Müller, G., Röglinger, M.: Business and Information Systems Engineering: A Complementary Approach to Information Systems - What We Can Learn from the Past and May Conclude from Present Reflection on the Future. *Journal of the Association for Information Systems* (2012)
6. Buhl, H.U., Fridgen, G., König, W., Röglinger, M., Wagner, C.: Where's the Competitive Advantage in Strategic Information Systems Research? – Making the Case for Boundary-spanning Research Based on the German Business and Information Systems Engineering Tradition. *Journal of Strategic Information Systems* (2012)
7. Baskerville, R., Lyytinen, K., Sambamurthy, V., Straub, D.: A response to the design-oriented information systems research memorandum. *European Journal of Information Systems* 20(1), 11–15 (2011)
8. Gill, G., Bhattacharjee, A.: Whom are we informing? Issues and recommendations for MIS research from an informing sciences perspective. *Management Information Systems Quarterly* 33(2), 217–235 (2009)
9. Lyytinen, K., Baskerville, R., Iivari, J., Te'eni, D.: Why the old world cannot publish? Overcoming challenges in publishing high-impact IS research. *European Journal of Information Systems* 16(4), 317–326 (2007)
10. Hasenkamp, U., Stahlknecht, P.: Wirtschaftsinformatik - Evolution of the Discipline as Reflected by its Journal. *Business & Information Systems Engineering* 1(1), 14–24 (2009)
11. Gregor, S.: The Nature of Theory in Information Systems. *MIS Quarterly* 30(3), 611–642 (2006)
12. Lee, A.: Retrospect and prospect: information systems research in the last and next 25 years. *Journal of Information Technology* 25(4), 336–348 (2010)
13. Mertens, P.: Die Zielfunktion des Universitätslehrers der Wirtschaftsinformatik - Setzen wir falsche Anreize? In: *Proceedings of the 11th International Conference on Business and Information Systems Engineering*, vol. 2, pp. 1167–1175 (February 2011)
14. Agarwal, R., Lucas, H.: The Information Systems Identity Crisis: Focusing on High-Visibility and High-Impact Research. *MIS Quarterly* 29(3), 381–398 (2005)
15. Winter, R.: Interview with Alan R. Hevner on “Design Science”. *Business & Information Systems Engineering* 1(1), 126–129 (2009)
16. Hirschheim, R., Klein, H.: Crisis in the IS Field? A Critical Reflection on the State of the Discipline. *Journal of the Association for Information Systems* 4(10), 237–293 (2003)
17. Winter, R.: Interview with Jay F. Nunamaker, Jr. on Toward a Broader Vision of IS Research. *Business & Information Systems Engineering* 2(5), 321–329 (2010)



18. Firth, D., Lawrence, C., Looney, C.A.: Addressing the IS Enrollment Crisis: A 12-step Program to Bring about Change through the Introductory IS Course. *Communications of the Association for Information Systems* 23(1), 2–36 (2008)
19. Hirschheim, R., Newman, M.: Houston, we've had a problem...offshoring, IS employment and the IS discipline: perception is not reality. *Journal of Information Technology* 25(4), 358–372 (2010)
20. Navarro, P.: The MBA core curricula of top-ranked U.S. business schools: a study in failure. *The Academy of Management Learning and Education* 7(1), 108–123 (2008)
21. Sabherwal, R.: Declining IS enrollments: a broader view of causes and strategies – a response to 'Houston, we've had a problem ... offshoring, IS employment and the IS discipline: perception is not reality'. *JIT* 25(4), 382–384 (2010)
22. Gill, G., Bhattacharjee, A.: Fashion Waves versus Informing? Response to Baskerville and Myers. *MIS Quarterly* 33(4), 667–671 (2009)
23. King, J.L., Myers, M.D., Rivard, S., Saunders, C., Weber, R.: What Do We Like About the IS Field? *Communications of the AIS* 26(1), 441–450 (2010)
24. Klein, H., Rowe, F.: Marshaling the professional experience of doctoral student: A contribution to the practical relevance debate. *MIS Quarterly* 32(4), 675–686 (2008)
25. Myers, M.D., Baskerville, R.L.: Commentary on Gill and Bhattacharjee: Is There an Informing Crisis? *MIS Quarterly* 33(4), 663–665 (2009)
26. Somers, M.: Using the theory of the professions to understand the IS identity crisis. *European Journal of Information Systems* (19), 382–288 (2010)
27. Taylor, H., Dillon, S., van Wingen, M.: Focus and Diversity in Information Systems Research: Meeting the dual demand of a healthy applied discipline. *MIS Quarterly* 34(4), 647–667 (2010)
28. Ghoshal, S.: Bad management theories are destroying good management practices. *Academy of Management Learning & Education* 4(1), 75–91 (2005)
29. Kagermann, H.: Innovation und Nutzen durch Offenheit. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung* 62, 673–674 (2010)

# Mining Constraints for Artful Processes<sup>\*</sup>

Claudio Di Ciccio and Massimo Mecella

SAPIENZA – Università di Roma  
Dipartimento di Ingegneria Informatica,  
Automatica e Gestionale ANTONIO RUBERTI  
{cdc,mecella}@dis.uniroma1.it

**Abstract.** Artful processes are informal processes typically carried out by those people whose work is mental rather than physical (managers, professors, researchers, engineers, etc.), the so called “knowledge workers”. MAILOFMINE is a tool, the aim of which is to automatically build, on top of a collection of email messages, a set of workflow models that represent the artful processes laying behind the knowledge workers activities. After an outline of the approach and the tool, this paper focuses on the mining algorithm, able to efficiently compute the set of constraints describing the artful process. Finally, an experimental evaluation of it is reported.

**Keywords:** process mining, artful process, declarative workflow, email.

## 1 Introduction

For a long time, formal business processes (e.g., the ones of public administrations, of insurance/financial institutions, etc.) have been the main subject of workflow related research. Informal processes, a.k.a. “artful processes”, are conversely carried out by those people whose work is mental rather than physical (managers, professors, researchers, engineers, etc.), the so called “knowledge workers” [30]. In contrast to business processes that are formal and standardized, often informal processes are not even written down, let alone defined formally, and can vary from person to person even when those involved are pursuing the same objective. Knowledge workers create informal processes “on the fly” to cope with many of the situations that arise in their daily work. Though informal processes are frequently repeated, they are not exactly reproducible even by their originators – since they are not written down – and can not be easily shared either. Their outcomes and their information exchanges are done very often by means of email conversations, which are a fast, reliable, permanent way of keeping track of the activities that they fulfill.

Understanding artful processes involving knowledge workers is becoming crucial in many scenarios. Here we mention some of them:

---

<sup>\*</sup> This work has been partly supported by SAPIENZA – Università di Roma through the grants FARI 2010 and TESTMED, and by the EU Commission through the FP7 project Smart Vortex. The authors would like also to thank Monica Scannapieco and Diego Zardetto for useful insights and discussions.

- *personal information management (PIM)*, i.e., how to organize one’s own activities, contacts, etc. through the use of software on laptops and smart devices (iPhones/iPads, smartphones, tablets). Here, inferring artful processes in which a person is involved allows the system to be proactive and thus drive the user through its own tasks (on the basis of the past) [11,30];
- *information warfare*, especially in supporting anti-crime intelligence agencies: let us suppose that a government bureau is able to access the email account of a suspected person. People planning a crime or an act out of law are used to speak a language of their own to express duties and next moves, where meanings may not match with the common sense. Thus, a system should build the processes that lay behind their communications anyway, exposing the activities and the role of the actors. At that point, translating the sense of misused words becomes an easier task for investigators, and allows inferring the criminal activities of the suspected person(s);
- *enterprise engineering*: in design and engineering, it is important to preserve more than just the actual documents making up the product data. Preserving the “soft knowledge” of the overall process (the so-called product life-cycle) is of critical importance for knowledge-heavy industries. Hence, the idea here is to take to the future not only the designs, but also the knowledge about processes, decision making, and people involved [20,28].

The objective of the MAILOFMINE approach, firstly introduced in [16], is to automatically discover, on top of a collection of email messages, a set of workflow models that represent the artful processes laying behind the knowledge workers’ activities. In [14], we describe our ideas on how to effectively show the users such models. In this paper, we outline the general approach of the mining algorithm, able to infer the constraints that specify the workflow altogether, out of the given execution traces. Then we present some experiments, showing the validity and efficiency of the technique.

The work presented here is related to the so called *process mining*, a.k.a. *workflow mining* [3], that is the set of techniques allowing the extraction of structured process descriptions from a set of recorded real executions (stored in the *event logs*). ProM [4] is one of the most used plug-in based software environment for implementing workflow mining techniques. Most of the mainstream process mining tools model processes as Workflow Nets (WFNs – see [2]), explicitly designed to represent the control-flow dimension of a workflow. From [7] onwards, many techniques have been proposed, in order to address specific issues: pure algorithmic (e.g.,  $\alpha$  algorithm [6] and its evolution  $\alpha^{++}$  [32]), heuristic (e.g., [31]), genetic (e.g., [22]). Indeed, heuristic and genetic algorithms have been introduced to cope with noise, that the pure algorithmic techniques were not able to manage. A very smart extension to the previous research work has been recently achieved by the two-steps algorithm proposed in [1].

The need for flexibility in the definition of processes leads to an alternative to the classical “imperative”: the “declarative” approach. Rather than using a procedural language for expressing the allowed sequences of activities, it is based on the description of workflows through the usage of constraints: the idea

is that every task can be performed, except what does not respect them. Such constraints, in DecSerFlow [5] and ConDec [24] (now named Declare [25,26]), are formulations of Linear Temporal Logic and have a graphical representation as well. [21] outlines an algorithm for mining Declare processes, implemented in ProM. The technique is based on [18,19,26], for the translation of Declare constraints into automata, and [33], for the optimization of such task.

[12] described the usage of inductive logic programming techniques to mine models expressed as a SCIFF [9] theory, finally translated to the ConDec notation.

The technique introduced in this paper differs from both [12] and [21] in that it does not directly verify the candidate constraints over the whole set of traces in input. It prepares an ad-hoc knowledge base of its own, so to further analyze the response to specific queries.

We believe that the declaration of collaborative workflows constraints can be expressed by means of regular expressions, rather than LTL formulae: regular expressions express finite languages (i.e., processes with finite traces, where the number of enacted tasks is limited). LTL formulae are thought to be used for verifying properties over semi-infinite runs instead. On the contrary, human processes have an end, other than a starting point. We envision the process schemes like grammars describing the language spoken by collaborative entities in terms of activities, thus being more related to formal languages rather than temporal logic. Finally, we can exploit the scientific results from the literature in the pattern matching (e.g., [17,8,29]).

The remainder of this paper is organized as follows: Section 2 outlines the overall approach of MAILOFMINE, then Section 3 describes the process model we adopt in our mining approach. Section 4 describes the relevant features of the mining technique; the interested reader can find more details, formal definitions, proofs and technicalities in [15]. Section 5 presents an extensive validation of the technique, and finally Section 6 draws some concluding remarks, outlining future activities.

## 2 The MAILOFMINE Approach

The MAILOFMINE approach (and the tool we are currently developing) adopts a modular architecture, the components of which allow to incrementally refine the mining process; before briefly presenting it, we need to introduce some basic concepts needed in the following. The details of the architecture, as well as the formal definitions of concepts, are presented in [16].

An *actor* is the subject directly or indirectly taking part in the progress of a work. A *task* is an elementary unit of work. Each task is connected to (i) its expected *duration*, (ii) zero or more *outcomes*, (iii) one or more *actors*. An *activity* is a collection of tasks or (recursively) other activities. A *key part* is each *unique* piece of text belonging to the email messages exchanged in a communication trace. For instance, signatures put in the footer by the senders of the email messages, quotations of previous email messages used in replies, etc., are all examples of redundant information that may appear more than once in a thread: they have to be filtered out by means of the key part concept, then. On the other hand, any piece of text, previously unread during the automated analysis of the thread, is interpreted as a key part, since it is supposed to add some information to the discussion (and

thus, to the underlying enacted process). An *indicium* is any communication trace, or part of it, attesting the likely execution of a task (task *indicium*) or an activity (activity *indicium*). In other words, it is any text where you can find an evidence that a task (or an activity) has been performed. A *process scheme* (or *process* for short) is a semi-structured set of activities, where the semi-structuring connective tissue is represented by the set of constraints stating the interleaving rules among activities or tasks. Constraints do not force the process instance to follow a tight sequence, but rather leave it the flexibility to follow different paths, while performing the execution, though respecting a set of rules that avoid illegal or non-consistent states. We argue that each constraint is expressible through regular grammars. Regular grammars are recognizable through Finite State Automata (FSA) [13] (either deterministic or non-deterministic [27]). The FSA recognizing the correct traces for processes (i.e., accepting the valid strings) is the intersection of all the FSAs composing the set of constraints.

Initially, we need to extract email messages from the given archive(s), so to analyze their raw data, including senders, recipients, headers, base-64-encoded attachments, etc., and extract the relevant information, in order to further elaborate and store it into a database. On top of the latter, all the subsequent steps are carried out. The first of them is the clustering of retrieved messages into extended communication threads, i.e., flows of messages which are related to each other. The considered technique, which has been described in [16], is based not only on the Subject field (e.g., looking at “Fwd:” or “Re:” prefixes) or SMTP Header fields (e.g., reading the “In-Reply-To” field), but on the application of a more complex object matching decision method. Once the communication threads are recognized, we can assume them all as possible proofs (namely, *indicia*) of an enacted activity.

Afterwards, email messages are cleaned up from signatures and quotations citing some text already written in another message within the thread, by the usage of a combination of the techniques described in [10] and [23]. The key parts are the text remaining in bodies and subjects once such filtering operation has been performed. MAILOFMINE can thus build the activity *indicia* as the concatenation of all the key parts belonging to the messages of a communication thread. Then, the clustering algorithm is used again, this time to identify the matches between activity *indicia*. By taking into account the activities set and the key parts, the clustering algorithm checks for matching key parts: those are considered task *indicia*. At this point, MAILOFMINE starts searching for execution constraints between tasks. Presenting how this technique works, and validating it, is the aim of this paper; it will be detailed in Section 4.

### 3 The Process Model

As previously introduced, a *process scheme* (or *process* for short) is a semi-structured set of activities, where the semi-structuring connective tissue is represented by the set of constraints stating the interleaving rules among activities or tasks. Constraints do not force the tasks to follow a tight sequence, but rather leave them the flexibility to follow different paths, though respecting a set of rules that avoid illegal or non-consistent states in the execution.

Here, we adopt the Declare [21] taxonomy of constraints as the basic language for defining artful processes in a declarative way. But whereas in Declare constraints are translated into LTL formulas, we express each constraint through regular expressions, as discussed in Section 3 of [15].

The *Existence*( $m, a$ ) constraint imposes the  $a$  character to appear at least  $m$  times in the trace. The *Absence*( $n, a$ ) constraint holds if  $a$  occurs at most  $n - 1$  times in the trace. *Init*( $a$ ) makes each trace start with  $a$ . *RespondedExistence*( $a, b$ ) holds if, whenever  $a$  is read,  $b$  was already read or is going to be read (i.e., no matter if before or afterwards). Instead, *Response*( $a, b$ ) enforces it by forcing a  $b$  to appear after  $a$ , if  $a$  was read. *Precedence*( $a, b$ ) forces  $b$  to occur after  $a$  as well, but the condition to be verified is that  $b$  was read - namely, you can not have any  $b$  if you did not read an  $a$  before. *AlternateResponse*( $a, b$ ) and *AlternatePrecedence*( $a, b$ ) both strengthen respectively *Response*( $a, b$ ) and *Precedence*( $a, b$ ) by stating that *each*  $a$  ( $b$ ) must be followed (preceded) by at least one occurrence of  $b$  ( $a$ ). The “alternation” is in that you can not have two  $a$  ( $b$ ) in a row before  $b$  ( $a$ ). *ChainResponse*( $a, b$ ) and *ChainPrecedence*( $a, b$ ), in turn, specialize *AlternateResponse*( $a, b$ ) and *AlternatePrecedence*( $a, b$ ), both declaring that no other character can occur between  $a$  and  $b$ . The difference between the two is in that the former is verified for each occurrence of  $a$ , the latter for each occurrence of  $b$ . *CoExistence*( $a, b$ ) holds if both *RespondedExistence*( $a, b$ ) and *RespondedExistence*( $b, a$ ) hold. *Succession*( $a, b$ ) is valid if *Response*( $a, b$ ) and *Precedence*( $a, b$ ) are verified. The same holds with *AlternateSuccession*( $a, b$ ), equivalent to the conjunction of *AlternateResponse*( $a, b$ ) and *AlternatePrecedence*( $a, b$ ), and with *ChainSuccession*( $a, b$ ), with respect to *ChainResponse*( $a, b$ ) and *ChainPrecedence*( $a, b$ ). *NotChainSuccession*( $a, b$ ) expresses the impossibility for  $b$  to occur immediately after  $a$ . *NotSuccession*( $a, b$ ) generalizes the previous by imposing that, if  $a$  is read, no other  $b$  can be read until the end of the trace. *NotCoExistence*( $a, b$ ) is even more restrictive: if  $a$  appears, not any  $b$  can be in the same trace.

In addition to the above constraints, we introduced some new ones (cf. Section 3 of [15]). Taking inspiration from the relational data model cardinality constraints, we call (i) *Participation*( $a$ ) the *Existence*( $m, a$ ) constraint for  $m = 1$  and (ii) *Unique*( $a$ ) the *Absence*( $n, a$ ) constraint for  $n = 2$ , since the former states that  $a$  must appear at least once in the trace, whereas the latter causes  $a$  to occur no more than once. *End*( $a$ ) is the dual of *Init*( $a$ ), in the sense that it constrain each trace to end with  $a$ . Beware that *End*( $a$ ) would be clueless in a LTL interpretation, since LTL is thought to express temporal logic formulae over infinite traces. On the contrary, it makes perfectly sense to have a concluding task for a finite process, expressed by means of a regular automaton like the one underlying any regular expression.

### 3.1 An Example

Here we outline a brief example. Let us suppose to have an email archive, containing various process instances indicia, and to focus specifically on the

planning of a new meeting for a research project. We suppose to execute the overall MAILOFMINE technique, and we report the possible result, starting from the list of tasks in activities (Process Description [1](#)). Then we consider the tasks of the “Agenda” activity only.

---

### Process Description 1. Activities and tasks list

---

Activity:	$\langle Agenda \rangle$
Task:	p (“proposeAgenda”): Productive
Actors:	{ <i>You</i> : Contributor, <i>Community</i> : Spectator}
Duration:	4 dd.
Task:	r (“requestAgenda”): Clarifying
Actors:	{ <i>Participant</i> : Contributor, <i>Community</i> : Spectator}
Duration:	$\perp$
Task:	c (“commentAgenda”): Clarifying
Actors:	{ <i>Participant</i> : Contributor, <i>Community</i> : Spectator}
Duration:	$\perp$
Task:	n (“confirmAgenda”): Productive
Actors:	{ <i>You</i> : Contributor, <i>Community</i> : Spectator}
Duration:	2 dd.

---

We suppose that a final agenda will be committed (“confirmAgenda” – n) after that requests for a new proposal (“requestAgenda” – r), proposals themselves (“proposeAgenda” – p) and comments (“commentAgenda” – c) have been circulated.

Some terms used in the example, e.g., *You*, Contributor, Community, Productive, Clarifying, etc. refer to the classification MAILOFMINE operates on messages during the mining phases, and are explained in [16](#). For the purposes of this paper, the reader can skip them. The aforementioned tasks and activities are bound to the following constraints. We start with the *existence* constraints of Process Description [2](#), each focusing on a single task.

---

### Process Description 2. Existence constraints on the example tasks and activities

---

Activity:	$\langle Agenda \rangle$
Task:	“confirmAgenda”, n: <i>Participation</i> (n), <i>Unique</i> (n), <i>End</i> (n)

---

In Process Description [3](#) we report the *relation* constraints, namely holding between couple of tasks.

---

### Process Description 3. Relation constraints on the example tasks and activities

---

Activity:	$\langle Agenda \rangle$
	<i>response</i> (r, p)
	<i>respondedExistence</i> (c, p)
	<i>succession</i> (p, n)

---

## 4 The MINERful Algorithm

*MINERful* is the algorithm for mining declarative constraints out of activities' traces. The previous stages of the MAILOFMINE approach allow tasks to be abstracted like characters appearing over finite strings, which, in turn, represent process traces. Thus, MINERful works on collections of finite strings, actually. Therefore, it solves the more general problem of finding a specific set of regular patterns out of a number of strings. We will interchangeably use the terms "task" and "character", as well as "trace" and "string", then.

MINERful is based on the concept of *MINERfulKB*: it holds all of the useful information extracted from the given traces and tailored to the further discovery of constraints that possibly lay behind. The first step of MINERful is to synthesize such a matrix, actually, in order to easily mine the declarative model afterwards. The details of the structure of MINERfulKB, as well as the formal definitions related to it, are omitted for sake of space. They can be found in [15]. The MINERfulKB is composed by MINERful interplays and MINERful ownplays.

The *MINERful interplay* is referred to a couple of characters, the first considered as the *pivot*,  $\rho$ , and latter considered as the *searched*,  $\sigma$ . For sake of simplicity, examples will consider **a** as the pivot  $\rho$  and **b** as the searched  $\sigma$ , over an alphabet  $\Sigma = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ . The MINERful interplay consists of (i) a function  $\delta_{\rho,\sigma}$ , mapping a distance between  $\rho$  and  $\sigma$  to the number of cases they appeared at that distance in a string<sup>1</sup>, (ii) a scalar  $b_{\rho,\sigma}^{\rightarrow}$ , the *in-between onwards appearances* counter, and (iii) a scalar  $b_{\rho,\sigma}^{\leftarrow}$ , the *in-between backwards appearances* counter. As examples,  $\delta_{\rho,\sigma}(2) = 10$  means that we have the evidence of a searched  $\sigma$  appearing 2 characters after the pivot  $\rho$  (like in *cacbcc*) over 10 cases;  $b_{\rho,\sigma}^{\rightarrow} = 2$  means that the pivot  $\rho$  appeared 2 times between the preceding occurrence of  $\rho$  and the following first occurrence of the searched  $\sigma$  (as in the substring *ac-caacb*); and  $b_{\rho,\sigma}^{\leftarrow} = 3$  means that the pivot  $\rho$  appeared 3 times between the following occurrence of  $\rho$  and the preceding first occurrence of the searched  $\sigma$ , as in the substring *bcacaaca*. The MINERful interplay expresses a local view on two characters, measuring the *distances* and the *alternations* between the first and the second one. Thus, one should focus on substrings, starting from the pivot  $\rho$  and ending in the searched  $\sigma$  (or viceversa, in case of negative distances). Such substrings are not necessarily related to the *first* occurrence of  $\rho$  in the string: *any*  $\rho$  is the initial (final) character for a following (preceding) substring ending in (starting from)  $\sigma$ , each separately analyzed.

The *MINERful ownplay* refers to one character at a time (i.e., the pivot  $\rho$  over itself). It is composed by (i) a function  $\gamma_{\rho}$ , denoting the total amount of appearances per string, (ii) a scalar  $g_{\rho}^i$ , namely the number of strings where the

<sup>1</sup> A positive distance represents the number of characters between  $\rho$  and the *following*  $\sigma$ , whereas a negative one is for a *preceding*  $\sigma$ . With a slight abuse of notation, we consider  $\delta_{\rho,\sigma}(+\infty)$  and  $\delta_{\rho,\sigma}(-\infty)$  to denote the number of cases in which the searched  $\sigma$ , respectively, did not appear in a string *after* the pivot  $\rho$ , or did not appear in a string *before*  $\rho$ .  $\delta_{\rho,\sigma}(0)$  counts the number of cases in which the searched  $\sigma$  did not appear nor before neither after  $\rho$ .



pivot  $\rho$  appeared as the *initial* one, and (iii) a scalar  $g_\rho^l$ , namely the number of strings where the pivot  $\rho$  appeared as the *last* one.

For instance, suppose to have the string *aabbac*, being *a* the pivot. Then we have for  $\delta_{a,\cdot}$ ,  $b_{a,\cdot}^\rightarrow$  and  $b_{a,\cdot}^\leftarrow$  the values reported in Table 1, whereas  $\gamma_a = \left\{ \begin{array}{l} \langle 3, 1 \rangle \\ \langle x, 0 \rangle \forall x \in \mathbb{N} \setminus \{3\} \end{array} \right\}$ ,  $g_a^i = 1$ , and  $g_a^l = 0$ .

**Table 1.** Examples of statistical values stored in MINERfulKB

	$-\infty$	$\dots$	$-4$	$-3$	$-2$	$-1$	$0$	$+1$	$+2$	$+3$	$+4$	$+5$	$\dots$	$+\infty$	
$\delta_{a,a}$	0	0	1	1	0	1	0	1	0	1	1	0	0	0	$b_{a,a}^\rightarrow = 0$ ; $b_{a,a}^\leftarrow = 0$
$\delta_{a,b}$	2	0	0	0	1	1	0	1	2	1	0	0	0	0	$b_{a,b}^\rightarrow = 1$ ; $b_{a,b}^\leftarrow = 0$
$\delta_{a,c}$	3	0	0	0	0	0	0	1	0	0	1	1	0	0	$b_{a,c}^\rightarrow = 2$ ; $b_{a,c}^\leftarrow = 0$

The initial step of our technique is the construction of the MINERfulKB, having as the input a bag of strings  $T$  and an alphabet  $\Sigma_T$ . The algorithm in charge of it is called twice, one onwards, one backwards, i.e., reading the string from left to the right and viceversa (according to the Western Latin standard). The algorithm is designed to be completely on-line, i.e., it updates the MINERfulKB as new strings occur and as new characters in the string are read, with no need to go back on previous data in the end. The MINERfulKB is designed to be tailored to the further reasoning for constraints discovery. Thus this latter step becomes easier and faster, than analyzing it directly from the raw data (the bag of strings). At the same time, building the MINERfulKB had to be fast as well: moving the whole complexity to this step would take no advantage. The pseudo-code of the algorithm, as well as the details and running examples of its execution are presented in [15].

The next and final step of the technique is the identification of the constraints verified over the set of strings. These are directly expressible as predicates over the MINERfulKB, and therefore easily transposable into instructions for a verification algorithm. The details, as well as the definitions of all the constraints as predicates over the MINERfulKB, are presented in [15]. As an example, the *RespondedExistence*(*a*, *b*) constraint is expressed like the following:

$$RespondedExistence(a, b) \equiv \neg(\delta_{a,b}(0) > 0)$$

i.e., there is no string such that *b* was not read in if *a* was. A high-level description of the technique can be represented as in Algorithm 1.

---

**Algorithm 1.** The MINERful pseudo-code algorithm (bird-eye watching)

---

$\mathcal{K}_T \leftarrow \text{computeKBOnwards}(T, \Sigma_T)$   
 $\mathcal{K}_T \leftarrow \text{computeKBBackwards}(T, \Sigma_T)$   
 $\mathcal{B} \leftarrow \text{discoverConstraints}(\mathcal{K}_T, \Sigma_T, |T|)$

---

## 5 Validation and Experiments

A validation of the technique has been performed by using the example outlined in Section 3.1 as the starting point. Its aim is to show the efficiency of the mining technique, being the underlying algorithm polynomial wrt. the dimension of the input.

We tested the algorithm by varying the input in terms of alphabet size (different characters appearing in the strings), number of constraints valid over the strings, range of the number of characters per string. Starting with two tasks,  $n$  and  $p$ , and the related constraints, we made various experiments making the alphabet grow up to the original, thus including also  $r$  and  $c$ , plus an unconstrained task,  $e$ . The constraints ranged from the minimal set of four (*Unique*( $n$ ), *Participation*( $n$ ), *End*( $n$ ), *Succession*( $p, n$ )) to the maximal set of seven (including *Response*( $r, p$ ), *RespondedExistence*( $c, p$ ), *AlternatePrecedence*( $r, c$ )). The lengths of the strings ranged through intervals of  $\{[2 \dots 8], [3 \dots 12], [4 \dots 16], [5 \dots 20]\}$ . The number of strings ranged as the power of 10, from 100 up to 1000000 with an exponential step. For each of the preceding combination, 10 runs were performed, for a total amount of 4480 executions. The random strings were created by Xeger<sup>2</sup>, a Java open-source library for generating random text from regular expressions. All of the parameters, and not only constraints, were expressed in terms of regular expressions indeed, and their conjunction passed to the Xeger engine.

The machine was a Sony VAIO VGN-FE11H (an Intel Core Duo T2300 1.66 GHz (2 MB L2 cache) with 2 GB of DDR2 RAM at 667 Mhz), having Ubuntu Linux 10.04 as the operating system and Java JRE v1.6.

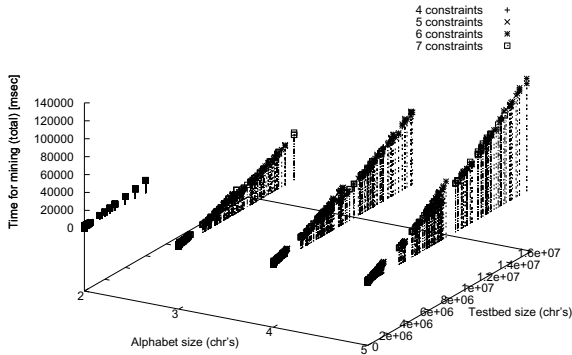
As the reader can see in Figure 1(a), the number of constraints does not affect the time taken by the algorithm to run – indeed, you are not able to distinguish between the curve designed by a group of points and another, where each group is related to a given number of constraints.

Figure 1(b) shows the fitting curves of the time taken by the algorithm to run, in comparison with the total amount of characters in input. It is a section of a parabola, confirming that the algorithm is quadratic w.r.t. the size of the strings.

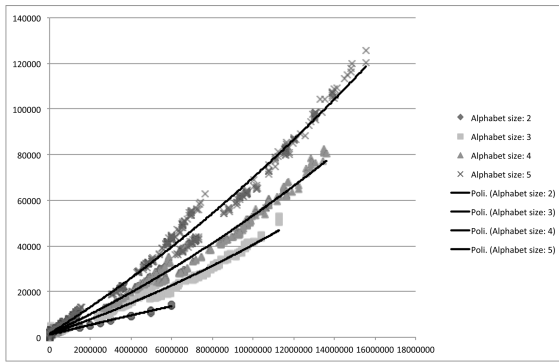
The algorithm is linear in the size of the collection of traces. In order to test this, we made a slightly different set-up: we fixed the number of constraints (7), the number of characters per string (10), and the alphabet length (5), whereas the number of strings ranged from 1000 to 12000 with a step of 1000. The result is depicted in Figure 1(c).

Furthermore, we report here by evidence that the time to build the MINERfulKB is the hardest task of the algorithm, with respect to the mining of constraints: in the worst case (1000000 strings, 7 constraints, 5 characters, length ranging from 5 to 20) the MINERfulKB was computed in 125.78 seconds whereas constraints were discovered in less than half a second (47 msec).

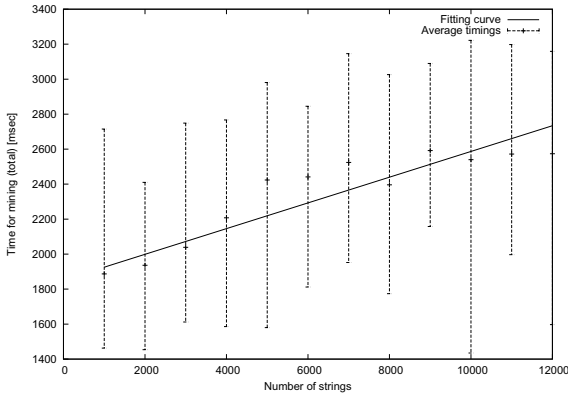
<sup>2</sup> <http://code.google.com/p/xeger/>



(a) Time needed for the execution, with respect to the testbed size and the alphabet size, setting the number of imposed constraints as parameter



(b) Execution time (ordinates), with respect to the testbed size (abscissae), setting the alphabet size as parameter



(c) Execution time, with respect to the number of strings, keeping fixed the size of strings, the alphabet length and the number of constraints

**Fig. 1.** Experimental results

## 6 Conclusions

As a concluding remark, we would like to highlight how the technique presented in this paper is only the last step of a complex approach, aimed at inferring artful processes from email messages; once that other techniques, out of the scope of this paper, allow us to consider email messages as strings over an alphabet of characters, the MINERful technique presented in this paper is able to infer which constraints are valid over such strings, thus inferring the process (described in a declarative way) that may lay behind them.

Further research activities are needed in order to refine and solve all of the used techniques and raised issues in MAILOFMINE. An extensive validation of the overall approach is planned as well. In this paper, we have shown that MINERful is a very efficient algorithm for mining constraints over strings. Throughout the experiments, we will be able to assess how much overfitting or underfitting MINERful is. We aim at validating it on a corpus of about 10 Gigabytes of email messages, derived from the activity of one of the authors in about 10 years of works in research projects, in order to infer common processes that partners adopted during software/deliverables' production. Then we will apply the approach to the field of collaborative activities of Open Source software development, reported by publicly available mailing lists.

## References

1. van der Aalst, W., Rubin, V., Verbeek, H., van Dongen, B., Kindler, E., Günther, C.: Process mining: a two-step approach to balance between underfitting and overfitting. *Software and Systems Modeling* 9, 87–111 (2010)
2. van der Aalst, W.M.P.: Verification of Workflow Nets. In: Azéma, P., Balbo, G. (eds.) ICATPN 1997. LNCS, vol. 1248, pp. 407–426. Springer, Heidelberg (1997)
3. van der Aalst, W.M.P.: Process mining: Discovery, conformance and enhancement of business processes. Springer (2011)
4. van der Aalst, W.M.P., van Dongen, B.F., Günther, C.W., Rozinat, A., Verbeek, E., Weijters, T.: ProM: The process mining toolkit. In: BPM 2009 Demos. *CEUR Workshop Proceedings*, vol. 489 (2009)
5. van der Aalst, W.M.P., Pesic, M.: DecSerFlow: Towards a Truly Declarative Service Flow Language. In: Bravetti, M., Núñez, M., Zavattaro, G. (eds.) WS-FM 2006. LNCS, vol. 4184, pp. 1–23. Springer, Heidelberg (2006)
6. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. *IEEE Trans. Knowl. Data Eng.* 16(9), 1128–1142 (2004)
7. Agrawal, R., Gunopulos, D., Leymann, F.: Mining Process Models from Workflow Logs. In: Schek, H.-J., Saltor, F., Ramos, I., Alonso, G. (eds.) EDBT 1998. LNCS, vol. 1377, pp. 469–483. Springer, Heidelberg (1998)
8. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE 1995, pp. 3–14 (1995)
9. Alberti, M., Chesani, F., Gavanelli, M., Lamma, E., Mello, P., Torroni, P.: Verifiable agent interaction in abductive logic programming: The SCIFF framework. *ACM Trans. Comput. Log.* 9(4) (2008)

10. de Carvalho, V.R., Cohen, W.W.: Learning to extract signature and reply lines from email. In: CEAS 2004 (2004)
11. Catarci, T., Dix, A., Katifori, A., Lepouras, G., Poggi, A.: Task-Centred Information Management. In: Thanos, C., Borri, F., Candela, L. (eds.) *Digital Libraries: Research and Development*. LNCS, vol. 4877, pp. 197–206. Springer, Heidelberg (2007)
12. Chesani, F., Lamma, E., Mello, P., Montali, M., Riguzzi, F., Storari, S.: Exploiting inductive logic programming techniques for declarative process mining. *T. Petri Nets and Other Models of Concurrency* 2, 278–295 (2009)
13. Chomsky, N., Miller, G.A.: Finite state languages. *Information and Control* 1(2), 91–112 (1958)
14. Di Ciccio, C., Mecella, M., Catarci, T.: Representing and Visualizing Mined Artful Processes in MailOfMine. In: Holzinger, A., Simoncic, K.-M. (eds.) *USAB 2011*. LNCS, vol. 7058, pp. 83–94. Springer, Heidelberg (2011)
15. Di Ciccio, C., Mecella, M.: MINERful, a mining algorithm for declarative process constraints in MailOfMine. Tech. rep. SAPIENZA Università di Roma (2012), [http://ojs.uniroma1.it/index.php/DIS\\_TechnicalReports/issue/view/416](http://ojs.uniroma1.it/index.php/DIS_TechnicalReports/issue/view/416)
16. Di Ciccio, C., Mecella, M., Scannapieco, M., Zardetto, D., Catarci, T.: MailOfMine – Analyzing mail messages for mining artful collaborative processes. In: *SIMPDA 2011*, pp. 45–59 (2011)
17. Garofalakis, M.N., Rastogi, R., Shim, K.: SPIRIT: Sequential pattern mining with regular expression constraints. In: *VLDB 1999*, pp. 223–234 (1999)
18. Gerth, R., Peled, D., Vardi, M.Y., Wolper, P.: Simple on-the-fly automatic verification of linear temporal logic. In: *PSTV 1995*, pp. 3–18 (1995)
19. Giannakopoulou, D., Havelund, K.: Automata-based verification of temporal properties on running programs. In: *ASE 2001*, pp. 412–416 (2001)
20. Heutelbeck, D.: Preservation of Enterprise Engineering Processes by Social Collaboration Software. In: Altmann, J., Baumöl, U., Krämer, B.J. (eds.) *Advances in Collective Intelligence 2011*. AISC, vol. 113, pp. 115–132. Springer, Heidelberg (2012)
21. Maggi, F.M., Mooij, A.J., van der Aalst, W.M.P.: User-guided discovery of declarative process models. In: *CIDM 2011*, pp. 192–199 (2011)
22. Medeiros, A.K., Weijters, A.J., Aalst, W.M.: Genetic process mining: an experimental evaluation. *Data Min. Knowl. Discov.* 14(2), 245–304 (2007)
23. Myers, E.W.: An O(ND) difference algorithm and its variations. *Algorithmica* 1(2), 251–266 (1986)
24. Pesic, M., van der Aalst, W.M.P.: A Declarative Approach for Flexible Business Processes Management. In: Eder, J., Dustdar, S. (eds.) *BPM Workshops 2006*. LNCS, vol. 4103, pp. 169–180. Springer, Heidelberg (2006)
25. Pesic, M., Schonenberg, H., van der Aalst, W.M.P.: Declare: Full support for loosely-structured processes. In: *EDOC 2007*, pp. 287–300 (2007)
26. Pesic, M., Schonenberg, M.H., Sidorova, N., van der Aalst, W.M.P.: Constraint-Based Workflow Models: Change Made Easy. In: Meersman, R., Tari, Z. (eds.) *OTM 2007, Part I*. LNCS, vol. 4803, pp. 77–94. Springer, Heidelberg (2007)
27. Rabin, M.O., Scott, D.: Finite automata and their decision problems. *IBM J. Res. Dev.* 3, 114–125 (1959)
28. Smart Vortex Consortium: Smart Vortex – Management and analysis of massive data streams to support large-scale collaborative engineering projects. FP7 IP Project, <http://www.smartvortex.eu/>

29. Srikant, R., Agrawal, R.: Mining Sequential Patterns: Generalizations and Performance Improvements. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) EDBT 1996. LNCS, vol. 1057, pp. 3–17. Springer, Heidelberg (1996)
30. Warren, P., Kings, N., Thurlow, I., Davies, J., Buerger, T., Simperl, E., Ruiz, C., Gomez-Perez, J.M., Ermolayev, V., Ghani, R., Tilly, M., Bösser, T., Imtiaz, A.: Improving knowledge worker productivity - the Active integrated approach. *BT Technology Journal* 26(2), 165–176 (2009)
31. Weijters, A., van der Aalst, W.: Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering* 10 (2001, 2003)
32. Wen, L., van der Aalst, W.M.P., Wang, J., Sun, J.: Mining process models with non-free-choice constructs. *Data Min. Knowl. Discov.* 15(2), 145–180 (2007)
33. Westergaard, M.: Better Algorithms for Analyzing and Enacting Declarative Workflow Languages Using LTL. In: Rinderle-Ma, S., Toumani, F., Wolf, K. (eds.) *BPM* 2011. LNCS, vol. 6896, pp. 83–98. Springer, Heidelberg (2011)

# Creating Sound and Reversible Configurable Process Models Using CoSeNets\*

Dennis M.M. Schunselaar, Eric Verbeek,  
Wil M.P. van der Aalst, and Hajo A. Reijers

Eindhoven University of Technology,  
P.O. Box 513, 5600 MB, Eindhoven, The Netherlands  
{d.m.m.schunselaar,h.m.w.verbeek,w.m.p.v.d.aalst,h.a.reijers}@tue.nl

**Abstract.** All Dutch municipalities offer the same range of services, and the processes delivering these services are quite similar. Therefore, these municipalities can benefit from configurable process models. This requires the merging of existing process variants into configurable models. Unfortunately, existing merging techniques (1) allow for configurable process models which can be instantiated to unsound process models, and (2) are not always reversible, which means that not all original models can be obtained by instantiation of the configurable process model. In this paper, we propose to capture the control-flow of a process by a CoSeNet: a configurable, tree-like representation of the process model, which is sound by construction, and we describe how to merge two CoSeNets into another CoSeNet such that the merge is reversible. Initial experiments show that this approach does not influence complexity significantly, i.e. it results in similar complexities for the configurable process model compared to existing techniques, while it guarantees soundness and reversibility.

## 1 Introduction

Within the CoSeLoG project, we have 10 municipalities involved offering essentially the same set of services. The process models supporting these services are very similar, due to legislation and standardisation, but are different, due to *couleur locale* and local decision making. The goal of the project is to support the different process models via configurable process models.

Configurable process models are process models with configuration options. The user has the possibility to configure the configurable process model by making configuration choices for options. These configurations are used to deduce the process models from the configurable process model (instantiation), by taking the different choices for the configuration options into account. Obtaining a configurable process model can be done via the merger of process models. Merging a set of process models should be such that the behaviour of a configurable process model is (an over-approximation of) the union of allowed behaviour from the different process models.

---

\* This research has been carried out as part of the Configurable Services for Local Governments (CoSeLoG) project (<http://www.win.tue.nl/coselog/>).

We require that every instantiation of the configurable process model yields a sound process model [1]. Furthermore, to support the different municipalities, we want that the configurable process models are reversible [2], i.e. the models used for obtaining the configurable model should be instantiations of the configurable model. Both requirements should not impact the complexity of the resulting configurable process model significantly in comparison to the state-of-the-art. Applying existing techniques to obtain configurable process models from the CoSeLoG process models resulted in configurable process models which can be instantiated to unsound process models, and some techniques are not reversible.

To counter the aforementioned problems with existing techniques, and adhering to the requirements, we propose to capture the process models in tree-like representations of block-structured process models, which are sound by construction (see [3] for a comparison between block-structured and graph-structured process models). Since there is a straightforward transformation from this tree-like representation to, for instance, Petri nets, we can use the classical notion of soundness.

The configurable variant of these process models is captured in CoSeNets, a tree-like representation of configurable process models. Instantiating a CoSeNet always yields sound process models. Furthermore, the merger of two CoSeNets is always reversible. In order to show that the complexity is not significantly influenced, an empirical comparison is presented between our approach and existing techniques. This empirical evaluation shows that the complexity is comparable to, or lower than existing techniques, using a subset of the processes used in [4].

This paper is structured as follows: Sect. 2 lists the relevant related work. In Sect. 3, the CoSeNets are explained. We explain the merger of CoSeNets in Sect. 4. In Sect. 5, we show a comparison between our approach and existing approaches. Finally, Sect. 6 contains the conclusions and future work. For an extended version of this paper with more technical details, we refer the reader to [5].

## 2 Related Work

A number of configurable process modelling languages has been proposed. These modelling languages can be subdivided into two categories, i.e. *imperative* and *declarative*. Declarative languages constrain the allowed behaviour, i.e. everything is allowed unless stated otherwise. An example of a configurable declarative process modelling language is *Configurable Declare* [6]. Imperative languages are the opposite of declarative languages; imperative languages specify the allowed behaviour, i.e. nothing is allowed unless stated otherwise. *C-SAP WebFlow*, *C-BPEL*, *C-YAWL* [7], and *C-EPC* [8] are examples of imperative configurable process modelling languages. These configurable modelling languages do not always have to yield sound process models when being instantiated [9,10].

A number of merging techniques have been defined in literature. Gottschalk [7] elaborates on the merger of process models into a single configurable process



**Table 1.** Comparing the different merging techniques, note that the soundness property of Gottschalk [7] is after some postprocessing

Approach	Gottschalk [7]	Li et al. [11]	Mendling et al. [12]	La Rosa et al. [2]	Sun et al. [14]
Soundness	✓	✗	✗	✗	✗
Reversibility	✓	✗	✗	✓	✗

model (e.g. EPCs). All requirements are met, however, one has to perform a postprocessing step to transform the process model into a sound process model. Furthermore, Gottschalk allows for all possible different instantiations of the configurable process model, i.e. certain parts can be blocked even if it was not observed in any of the input models. Although this does not limit the reversibility, it is noteworthy from an application point of view since it will become harder in this way to identify which parts are commonalities and which parts are variabilities.

Li et al. [11] present an approach for creating a new reference model based on models mined from a log. These models represent the different variations of a reference model. Li et al. only consider “AND” and “XOR” operators. Furthermore, they seek to minimise a distance measure between the reference model and the mined models, where distance is defined as the number of insertions, deletions, and moving activities within the process model. By allowing the insertion of activities, the reference model cannot be configured to obtain the input models. Note that in our approach it is not allowed to insert activities.

In the paper by Mendling et al. [12], an approach is presented to merge the different views on a process model. This approach does not yield configurable models, i.e. the output is an EPC and not a C-EPC. Furthermore, soundness is not guaranteed by this particular approach.

La Rosa et al. [2] present an approach for merging a set of process models into a single configurable process model. With the use of AProMoRe [13] they are even able to merge different formalisms into a single configurable process model. Although this approach allows for the deduction of the input models, it does not guarantee the deduction of sound process models.

Sun et al. [14] focus on merging block-structured process models. They, however, provide a merge for disjoint process fragments, while in this paper we focus on the merger of variants of the *same* process. Due to the merger of disjoint process fragments, some activities are marked as redundant and removed from the resulting model, which we consider undesirable as it does not allow for the deduction of the input models. Finally, their approach does not allow for configurations, i.e. the resulting model is a non-configurable process model.

Table 1 lists the different merging approaches and their adherence to our requirements.

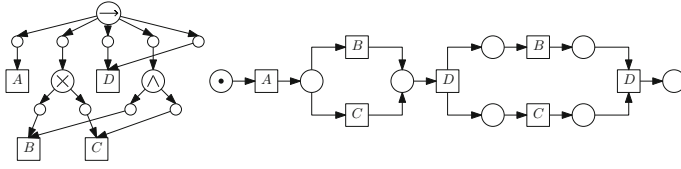


Fig. 1. A process model with the corresponding Petri net

### 3 CoSeNet and Metrics

Here, we introduce our new representations of process models and configurable process models, i.e. process models and CoSeNets. Furthermore, the metrics used for the experimental evaluation are elaborated on.

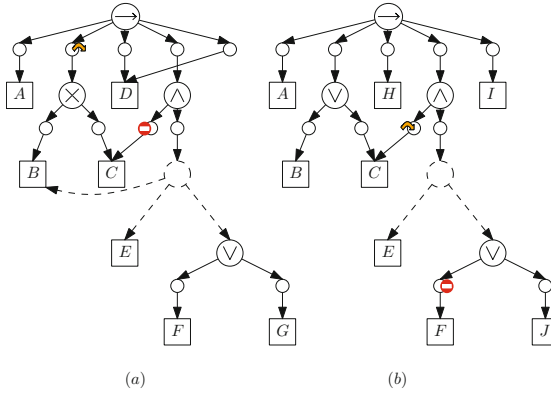
#### 3.1 Process Model

For our purposes, a process model is captured as a tree-like representation of a block-structured process model. However, we allow for sharing subprocesses, i.e. some subtrees might have multiple incoming edges to support reuse. Therefore, we use Directed Acyclic Graphs (DAGs) to represent process models.

Fig. 1 depicts a process model with its corresponding Petri net. The top node in the process model denotes the root and is a SEQ node (sequence,  $\rightarrow$ ). The SEQ node has 5 children, i.e. three activity nodes ( $A$  and twice  $D$ ) and two operator nodes: XOR ( $\times$ ) and AND ( $\wedge$ ). Although the context in which  $D$  is executed might be different for both  $D$ 's, from a control-flow perspective we assume they are the same  $D$ . A SEQ node executes its children in the order in which they occur, thus  $A$  is the first and the second  $D$  is the last. The XOR node denotes an exclusive choice between any of its children, and the AND node denotes the parallel execution of its children. Apart from the SEQ, XOR, and AND, we currently support the OR node (one can execute any number of children but at least one) and the DEF node (deferred choice, i.e. the choice based on events instead of data). Furthermore, we support LOOP nodes, which have three children: a *do* child, a *redo* child, and an *exit* child. The do child is the root of the subgraph representing the body of the loop. After having executed the do child, there is a choice to either execute the redo child and, afterwards, execute the do child again, or to execute the exit child and exit the loop construct. The choice between the redo child and the exit child can be either exclusive (LOOPXOR) or deferred (LOOPDEF). For the sake of brevity, we use LOOP as a shorthand for both LOOPXOR and LOOPDEF.

#### 3.2 CoSeNet

A CoSeNet is an extension of the process model we discussed hitherto. Fig. 2 depicts (a) the process model from Fig. 1 extended with some extra annotations and nodes, and (b) a second CoSeNet which we want to merge with (a). The stop



**Fig. 2.** Two CoSeNets we want to merge

sign denotes that this branch can be *blocked*, i.e. the execution of this subgraph can be prevented. *Hiding* a particular branch, i.e. substituting the branch by a silent transition ( $\tau$ ), is denoted with the orange arrow. Finally, we have added a *placeholder* node (dashed circle) to offer the user to select a subgraph to replace this placeholder node. In this example, the user can select to substitute the placeholder node by the activity node  $B$ ,  $E$ , or by the subgraph rooted at the OR node. A placeholder node is used instead of an exclusive choice to select a subgraph. This to shift the moment of choice from run-time to configuration-time. Fig. 1 can be obtained from (a) by not blocking the blockable branch, not hiding the hideable branch, and replacing the placeholder node by activity node  $B$ .

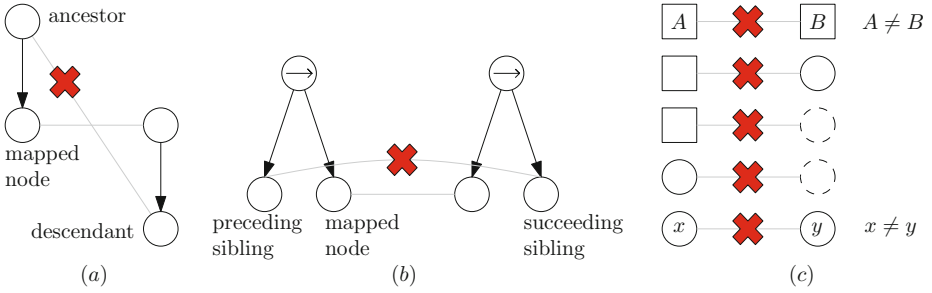
### 3.3 Metrics

In order to compare our approach with existing approaches w.r.t. complexity (specifically, La Rosa et al. [2] and Gottschalk [7]), we use the complexity metrics used in [4] since this is a continuation of that work. We elaborate briefly on the used metrics. For a more complete discussion, we refer the reader to [4,15,16,17].

*Control-Flow Complexity (CFC)* [16] computes for every operator a weight based on the number of outgoing edges. The CFC for a process model is the summation of the weights for the individual operators in that process model.

The *density* of a process model [15] is defined as the amount of edges in the model divided by the total amount of edges possible in that model.

With the *Cross-Connectivity (CC)* metric [17], one first computes the weight of the different nodes (connectors and tasks) in the process model (based on the amount of outgoing edges). Afterwards, the weight of the edge between two nodes is deduced from the weight of the nodes the edge is connected to. From this, the maximal weight for the paths between two nodes  $u$  and  $v$  is computed, where a path is a sequence of edges. Finally, the summation of paths with the



**Fig. 3.** Restrictions on a proper CoSeMap

largest weight between all pairs of nodes, is divided by the total amount of edges possible in a directed graph with  $N$  nodes (i.e.  $N \cdot (N - 1)$ ).

## 4 Merge

When merging two CoSeNets into a single CoSeNet it is important to know which nodes from the original CoSeNets may be merged into a single node. For this reason, we introduce the concept of a *node mapping* between both original CoSeNets, called a *CoSeMap*: Only if a node from one CoSeNet is mapped onto a node of the other CoSeNet, then these nodes may be merged into a single node. For the sake of simplicity, we assume a one-to-one correspondence between nodes, and a CoSeMap corresponds to an injective function to and from the nodes in both CoSeNets (for the sake of convenience, we assume a CoSeMap to be symmetrical).

As a CoSeNet corresponds to a DAG, the CoSeNet that results from a merge may not contain any cycles. Cycles may appear in the resulting CoSeNet if an ancestor node of a mapped node in one CoSeNet is mapped onto a descendant node of the node that the mapped node is mapped onto in the other CoSeNet (see Fig. 3(a)). Second, for a resulting sequence node a correct ordering of its children should be feasible, which is impossible if any preceding sibling of a mapped child in one CoSeNet is mapped onto a succeeding sibling of the node the mapped child is mapped onto in the other CoSeNet (see Fig. 3(b)). It makes no sense to map a node from one type (activity, operator, placeholder) to a node of another type, or to map an activity node to another activity node with a different label (see Fig. 3(c)). Finally, we only allow the mapping of LOOP nodes if and only if all nodes related to the LOOP nodes are mapped. I.e. the root of the *do* subgraph, *redo* subgraph, and the root of the *exit* subgraph. For these reasons, we require a CoSeMap to be *proper*:

1. No ancestor node is mapped onto a descendant node;
2. No preceding child is mapped onto a succeeding child for any sequence node;

3. All nodes are mapped onto nodes from the same type and with the same label (in case they are activity nodes);
4. Loop nodes are mapped onto each other if all children are mapped in order.

Given a proper CoSeMap between them, two CoSeNets can be merged in a straightforward way, which is the *CoSeMerge*:

1. All nodes from the first CoSeNet are added to the resulting CoSeNet;
2. All unmapped nodes from the second CoSeNet are added;
3. All edges from the first CoSeNet are added;
4. All edges that involve some unmapped node from the second CoSeNet are added, but only after any mapped node is replaced by the node it is mapped onto from the first CoSeNet;
5. Configuration options from the second CoSeNet that involve only mapped nodes are added (if needed) to a corresponding branch from the first CoSeNet. Furthermore, some extra configuration options have to be added (see [5], for the exact details);
6. A new root node is added, which is a placeholder node with both root nodes as children. If both root nodes are mapped onto each other, the new root node will only have an edge to the root of the first CoSeNet.

Please note that edges from a SEQ/LOOP node are added in such a way that the order in which the children occur of both original nodes are taken into account. The *CoSeMerge* yields a CoSeNet containing the union of behaviour of the input CoSeNets.

When applying our *CoSeMerge* on any two CoSeNets (say,  $N_1$  and  $N_2$ ), it is easy to see that  $N_1$  can be instantiated from the resulting CoSeNet as all nodes and edges from  $N_1$  are present: After having removed all other nodes and edges, and after having removed the new root placeholder node and configuration options that were only present in  $N_2$ , the resulting graph is identical to  $N_1$ . For  $N_2$  this is a bit harder to see, as some nodes and edges from this net are not present. However, it is clear that only those nodes and edges are left out for which an alternative exists in  $N_1$ . Thus, after having removed all other nodes and edges, and the new root placeholder node and configuration options only present in  $N_1$ , the CoSeNet is identical to  $N_2$ . Hence, both  $N_1$  and  $N_2$  can be instantiated from the resulting CoSeNet, which means that the *CoSeMerge* is reversible.

In the remainder of this paper, we will describe two different ways to construct a proper CoSeMap, called an *activity* CoSeMap and an *extended* CoSeMap. An activity CoSeMap maps only activity nodes; it is computed by taking all pairs of activity nodes with the same label. Recall that we assume that a CoSeNet does not contain two activity nodes with the same label. If a CoSeNet contains two activity nodes with the same label, we can combine these into a single activity node. Fig. 4 shows the resulting CoSeNet after having merged both CoSeNets from Fig. 2 using an activity CoSeMap. An extended CoSeMap takes an activity

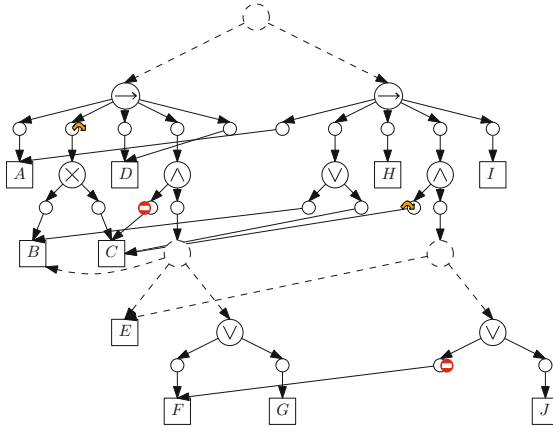


Fig. 4. Merging the models from Fig. 2 using an activity CoSeMap

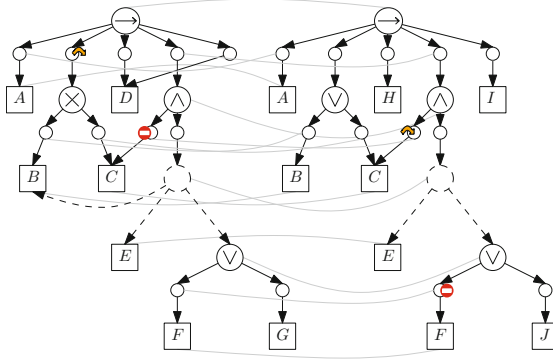


Fig. 5. An extended CoSeMap (horizontal gray lines) for merging the CoSeNETs from Fig. 2

CoSeMap as a starting point, but extends this CoSeMap with operator nodes and placeholder nodes (if possible) in such a way that the number of mapped nodes is maximised. The basic strategy for constructing the extended CoSeMap is to map one node onto another node (of the same type) if some child of the first node is mapped onto some child of the second node. As a result, the construction of this CoSeMap works in a bottom-up way, as initially we only have activity nodes that are mapped onto each other, which reside at the bottom of a CoSeNET. Note that it is possible that we have to choose between alternative ways to extend a current partial extended CoSeMap. Hence, given an activity map, multiple extended CoSeMaps may exist. If this is the case, we arbitrarily choose one of the alternatives. Fig. 5 shows a possible extended CoSeMap for merging both CoSeNETs from Fig. 2. Fig. 6 shows the result after having merged these CoSeNETs using this CoSeMap.

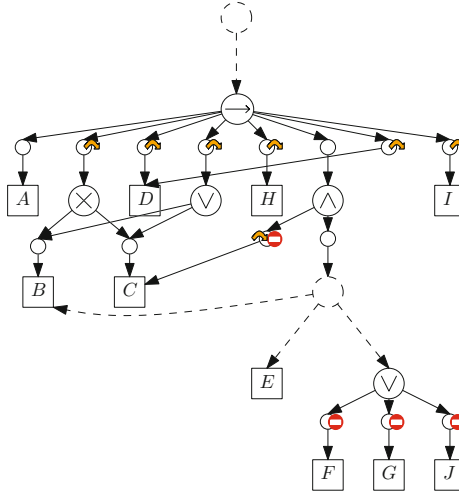


Fig. 6. Merging the models from Fig. 2 using the extended CoSeMap from Fig. 5

## 5 Experimental Evaluation

In the previous sections, we have argued that CoSeNets can only be instantiated to sound process models. Furthermore, the CoSeMerge ensures that the merged CoSeNet is reversible. In this section, we want to show that these attractive properties of soundness and reversibility do not incur a penalty on the complexity of the configurable process model.

The implementation of the construction of an activity CoSeMap, the construction of an extended CoSeMap, and the CoSeMerge have been implemented as ProM 6 plug-ins<sup>1</sup>. The construction of an activity CoSeMap is a straightforward implementation of string comparison on the activity labels. The construction of an extended CoSeMap is computed via linear programming. Linear programming is used since we have a maximisation problem, i.e. we desire that our extended CoSeMap is maximal. If our extended CoSeMap is maximal, it allows us to share as much subgraphs as possible, which reduces the duplication of subgraphs and hence the complexity of the configurable process model. After the merger, the CoSeNet is converted into a YAWL [18] model, which can be analysed in a repaired version of ProM 5.2<sup>2</sup>. Noteworthy facts of this conversion are that it treats a placeholder node as an XOR operator node, that it will reuse the YAWL fragment that corresponds to an operator node if possible, but that it will not reuse the YAWL fragment that corresponds to an activity node.

<sup>1</sup> <http://www.promtools.org/prom6/>

<sup>2</sup> <http://win.tue.nl/coselog/files/ProM-CoSeLoG-20110802.zip>, which corresponds to ProM 5.2 with some bugs in the algorithms to compute the various metrics have been fixed.

**Table 2.** The complexities of the models using different merging techniques

Municipality	GBA 1	GBA 2	GBA 3	MOR	WOZ
Mun <sub>A</sub>	5	21	11	42	12
Mun <sub>B</sub>	3	29	11	23	8
Mun <sub>C</sub>	2	38	28	29	14
Mun <sub>D</sub>	3	35	18	24	11
Mun <sub>E</sub>	6	25	26	25	25
Mun <sub>F</sub>	3	21	9	44	15
Mun <sub>G</sub>	5	21	9	20	15
Mun <sub>H</sub>	5	29	11	18	11
Mun <sub>I</sub>	3	41	9	28	11
Mun <sub>J</sub>	3	29	8	25	26
Act. CoSeMap	56	435	209	397	223
Ext. CoSeMap	39.8 ±4.3	126.3 ±10.6	172.9 ±23.2	262.4 ±18.9	134.1 ±13.6
La Rosa et al.	146.6 ±12.9	781.3 ±42.7	412.7 ±16.5	937.8 ±34.3	707.1 ±34.9
Gottschalk	80	317	210	-	335

We compare our approach to the approaches from Gottschalk [7] and La Rosa et al. [2]. As the process models used in [4] were not well-structured, we could not use these process models directly and had to modify them for our analysis.

Table 2 lists the values for the various complexity measures for the individual process models as well as for the merged process model. In case an approach yields a non-deterministic result, the average  $\mu$  and standard deviation  $\sigma$  of 10 results are listed as “ $\mu \pm \sigma$ ”. “Act. CoSeMap” and “Ext. CoSeMap” represent the merger with an activity CoSeMap and with an extended CoSeMap. The Synergia tool-set<sup>3</sup>, implements the merge by La Rosa et al. [2]. Finally, “Gottschalk” is the implementation of the EPC-merge by Gottschalk in ProM 5.2<sup>4</sup> [7].

From Table 2, one can see that the complexities of the different approaches vary significantly, i.e. ranging from roughly a factor 2 (GBA 1) up to roughly a factor 6 (GBA 2). Complexity metrics allow us to compare processes in a quantitative manner. In general, a higher complexity for an approach means a worse approach than an approach which yields models with a lower complexity (Occam’s Razor). See [15], for an extensive evaluation of the different metrics.

Placing the 10 models next to each other can be seen as a base case (which corresponds to the activity CoSeMap), as our conversion to YAWL will not reuse YAWL fragments that correspond to activity nodes, as mentioned earlier). We wish to avoid that the complexity of the configurable process model is significantly greater than the sum of the complexities of the individual models. In case of La Rosa et al. [2], they merge different connectors into a single connector which increases the CFC metric, especially with OR-operators which are exponential in the out-degree.

<sup>3</sup> <http://www.processconfiguration.com/>

<sup>4</sup> <http://promtools.org/prom5/>



The approach of Gottschalk [7] is in some cases comparable (w.r.t. complexity) to our approach (e.g. GBA 3), but yields in some cases significantly higher levels of complexity (e.g. GBA 2) or is not computable (e.g. MOR). In general, Gottschalk performs similar to the activity CoSeMap.

All in all, it can be concluded that, at least with this set of models, we achieve a comparable (or even lower) complexity w.r.t. existing techniques. Hence, it seems possible to obtain sound and reversible process models without paying a too expensive price for this, or actually no price at all.

## 6 Conclusion

Existing techniques allow for the instantiation of unsound process model from a configurable process model. Furthermore, some techniques are not reversible. Our solution successfully addresses both problems. Soundness is addressed by considering tree-like representations of process models and CoSeNets. The merge as we defined it in this paper takes care of the reversibility property. When merging CoSeNets into a single CoSeNet, the original CoSeNets are instantiations of the resulting CoSeNet.

Apart from defining our solution, we also applied our solution on the process models from the CoSeLoG project. This evaluation supports the view that the complexity of our approach is similar to/lower than the approach by Gottschalk and La Rosa et al. Thus, the guarantee of soundness and reversibility does not incur a penalty on the complexity of the configurable process models.

However, there is still room for improvement. This paper is, therefore, to be considered as a starting point. There are numerous ways in which we want to continue the development of this new approach. Amongst other in the following directions: this research has been started to support the processes of the municipalities. Therefore, we plan to extend our CoSeMerge and CoSeMaps, e.g. different process models have different granularity (see the work of Weidlich et al. [19]). Furthermore, as noted in the experimental evaluation, some quality dimensions have not (yet) been addressed. We intend to define these quality dimensions on our CoSeNets.

Finally, we would like to extend CoSeNets with resources and data, in order to fully support the processes of the municipalities.

**Acknowledgements.** We would like to thank Marcello La Rosa for his comments, which improved this paper significantly.

## References

1. van der Aalst, W.M.P.: Verification of Workflow Nets. In: Azéma, P., Balbo, G. (eds.) ICATPN 1997. LNCS, vol. 1248, pp. 407–426. Springer, Heidelberg (1997)
2. La Rosa, M., Dumas, M., Uba, R., Dijkman, R.M.: Business Process Model Merging: An Approach to Business Process Consolidation (in press, 2012)

3. Kopp, O., Martin, D., Wutke, D., Leymann, F.: The Difference Between Graph-Based and Block-Structured Business Process Modelling Languages. *Enterprise Modelling and Information Systems* 4(1), 3–13 (2009)
4. Vogelaar, J.J.C.L., Verbeek, H.M.W., Luka, B., van der Aalst, W.M.P.: Comparing Business Processes to Determine the Feasibility of Configurable Models: A Case Study. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) *BPM Workshops 2011, Part II*. LNBP, vol. 100, pp. 50–61. Springer, Heidelberg (2012)
5. Schunselaar, D.M.M., Verbeek, H.M.W., van der Aalst, W.M.P., Reijers, H.A.: Creating Sound and Reversible Configurable Processes Models using CoSeNets. Technical Report BPM Center Report BPM-11-21, BPMcenter.org (2011)
6. Schunselaar, D.M.M.: Configurable Declare. Master's thesis, Eindhoven University of Technology, The Netherlands (2011)
7. Gottschalk, F.: Configurable Process Models. PhD thesis, Eindhoven University of Technology, The Netherlands (2009)
8. Rosemann, M., van der Aalst, W.M.P.: A Configurable Reference Modelling Language. *Information Systems* 32(1), 1–23 (2007)
9. van der Aalst, W.M.P., Lohmann, N., La Rosa, M.: Ensuring Correctness During Process Configuration Via Partner Synthesis. *Information Systems* 37(6), 574–592 (2012)
10. van der Aalst, W., Lohmann, N., La Rosa, M., Xu, J.: Correctness Ensuring Process Configuration: An Approach Based on Partner Synthesis. In: Hull, R., Mendling, J., Tai, S. (eds.) *BPM 2010*. LNCS, vol. 6336, pp. 95–111. Springer, Heidelberg (2010)
11. Li, C., Reichert, M., Wombacher, A.: Discovering Reference Models by Mining Process Variants Using a Heuristic Approach. In: Dayal, U., Eder, J., Koehler, J., Reijers, H.A. (eds.) *BPM 2009*. LNCS, vol. 5701, pp. 344–362. Springer, Heidelberg (2009)
12. Mendling, J., Simon, C.: Business Process Design by View Integration. In: Eder, J., Dustdar, S. (eds.) *BPM Workshops 2006*. LNCS, vol. 4103, pp. 55–64. Springer, Heidelberg (2006)
13. La Rosa, M., Reijers, H.A., van der Aalst, W.M.P., Dijkman, R.M., Mendling, J., Dumas, M., García-Bañuelos, L.: APROMORE: An advanced process model repository. *Expert Systems with Applications* 38, 7029–7040 (2011)
14. Sun, S., Kumar, A., Yen, J.: Merging Workflows: A New Perspective on Connecting Business Processes. *Decision Support Systems* 42(2), 844–858 (2006)
15. Mendling, J.: Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness. Springer (2008)
16. Cardoso, J.: Control-flow Complexity Measurement of Processes and Weyuker's Properties. In: 6th International Enformatika Conference, Transactions on Enformatika, Systems Sciences and Engineering, vol. 8, pp. 213–218 (2005)
17. Vanderfeesten, I.T.P., Reijers, H.A., Mendling, J., van der Aalst, W.M.P., Cardoso, J.: On a Quest for Good Process Models: The Cross-Connectivity Metric. In: Bellahsene, Z., Léonard, M. (eds.) *CAiSE 2008*. LNCS, vol. 5074, pp. 480–494. Springer, Heidelberg (2008)
18. ter Hofstede, A.H.M., van der Aalst, W.M.P., Adams, M., Russell, N. (eds.): *Modern Business Process Automation: YAWL and its Support Environment*. Springer (2010)
19. Weidlich, M., Dijkman, R., Mendling, J.: The ICoP Framework: Identification of Correspondences between Process Models. In: Pernici, B. (ed.) *CAiSE 2010*. LNCS, vol. 6051, pp. 483–498. Springer, Heidelberg (2010)

# Semantics-Based Business Process Model Similarity

Bernhard G. Humm and Janina Fengel

Hochschule Darmstadt – University of Applied Sciences  
Haardtring 100  
64295 Darmstadt, Germany  
{bernhard.humm, janina.fengel}@h-da.de

**Abstract.** Business process modeling has become an accepted means for designing and describing business operations. As a result, comparing and aligning business process models within and between organizations is increasingly important. However, due to differing use of modeling languages and domain languages for labeling models and their elements, model comparison is a non-trivial task. Presently, it is to be performed manually. For easing this workload, we present a novel approach for determining semantic similarity in an automated manner, directed at supporting business analysis through semantic reasoning.

**Keywords:** Business process modeling, model comparison, semantic similarity, semantic reasoning.

## 1 Introduction

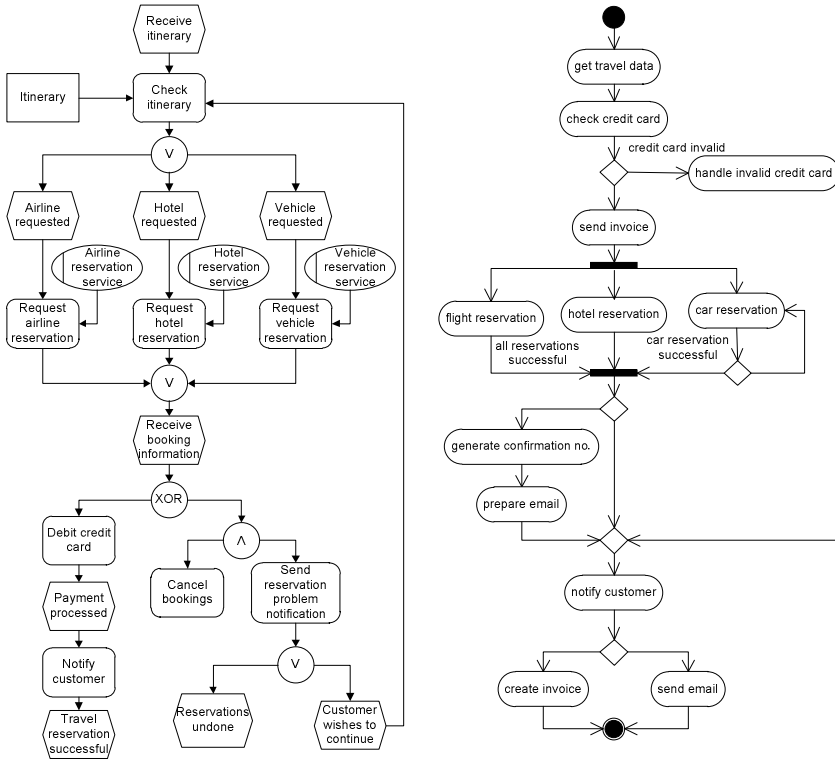
Businesses all over the world are faced with the challenge of having to flexibly react to change and to dynamically work with varying business partners. Continuous shaping and reshaping of business processes and the supporting or even enabling IT is a critical success factor for a business's competitiveness [1, 2]. For establishing electronic business, the underlying processes, required information and subsequent IT-support need to be described precisely. Over the past decades, business process modeling has become an accepted means for designing and describing business operations in enterprises within and across company boundaries. Such models describe interrelated business objects and business activities in a specific sequence, expressed in a certain modeling language with elements labeled in natural language. Thereby, major tasks in working with models are the analysis of these descriptions for the purpose of quality and compliance checking and detecting commonalities with models of different origin [3, 4]. This involves checking models with regard to the modeling languages and, most importantly, the domain language used for labeling the model elements. If the choice of words of the labels representing the business semantics is not dominated by rules, models are semantically heterogeneous, not only concerning their modeling language, but more importantly, concerning their domain language. This makes their comparison or integration a non-trivial task [5, 6].

As a result, differing types of models and dissimilarly applied business terminology prevent direct automated business process interactions without prior manual preparation efforts for resolving discrepancies. Typically, these challenges arise in all kinds of e-business integration projects, such as enterprise architecture, data and process integration scenarios [7]. Especially at the time of mergers and acquisitions and setting-up business collaborations where the models to be integrated originate from different independent sources, semantic analysis requires extensive intellectual efforts and time [8]. Therefore, in practice the problem presently to be tackled is the task of having to analyze hundreds of business process models manually. Models need to be compared regarding the intended meaning of their elements and their structure, whereas structural analysis cannot be performed until successful alignment of the domain language [9]. For easing this workload, automated support is deemed desirable by way of enabling (semi-)automatic alignment of process models concerning their semantic similarity. For supporting the analysis of models with regard to the intended meaning of language concepts present, the idea of applying semantic technologies in business process modeling has been suggested [10]. Semantic annotation of business process models has been proposed to allow for analyzing and comparing models [11, 6, 12]. As a complement to these efforts, we here present our method of analyzing and reasoning over business process models based on the domain facts contained.

We report on our research and continue with presenting a typical application scenario encountered, followed by a description of our method based on Semantic Web technologies and show its application. We conclude with an evaluation and discussion of our proposition together with a view onto related work and future research directions.

## 2 Motivating Example Application Scenario

To demonstrate our method, we show a typical application scenario for business process model integration. The motivating example is a situation occurring within enterprises at the time of outsourcing. It is a case of a travel agency implant where a company decided to hand over the booking of travel related services to an agency belonging to an independent business partner but installed on the premises. Fig. 1 shows two example models, an Event-Based Process Chain (EPC) called “Travel Reservation” (adapted from [13]) and an UML Activity Diagram called “Travel Booking” (adapted from [14]). The models describe a similar business operation, namely the booking of travel services, in different modeling languages and using different domain expressions as model element labels. As often encountered in real-world situations, the models contain errors and mismatches. Upon comparison, differences in the domain language usage can be detected, e.g., “Request airline reservation” corresponds to “flight reservation” and “vehicle reservation” corresponds to “car reservation”. Before resolving these differences, an analysis of the flow of activities depicted cannot be done.



**Fig. 1.** Two Example Business Process Models from the Travel Domain

In order to be able to reason over the models, they need to be analyzed and compared. On this basis, questions can be answered. The key question we focus on in this paper is the measure of *similarity* between two models. Thereby, similarity means the semantic correspondence of two models with respect to the business domain.

### 3 Business Process Model Similarity

There are numerous definitions of business process model in the literature [3, 15–17]. For our purposes, it is sufficient to postulate the following requirements for a business process model (or simply “model” if the context is clear): (a) it has a name; (b) it is defined in a business process modeling language, e.g., BPMN, EPC, or UML activity model; (c) it consists of labeled nodes and arcs that are optionally commented.

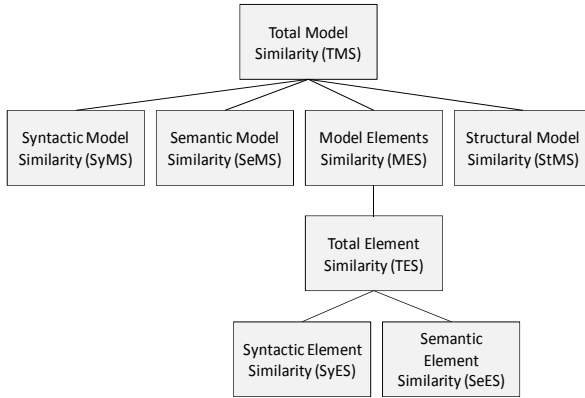
For comparing such models, we distinguish different dimensions of model similarity:

- *Syntactic similarity*: taking into account the types of modeling languages and language elements,
- *Semantic similarity*: taking into account the meaning of model and element labels and comments,
- *Structural similarity*: taking into account the model graph structure.

Syntactic and semantic similarity can both be applied to models as a whole as well as to individual model elements, e.g., nodes. We express the *measure of similarity* between two models as a numeric value between and including 0 and 1. The measure 1 denotes identity of two models, a smaller measure indicates a lower degree of similarity. This notion of similarity conforms to the notion of similarity measures according to [18].

### 3.1 Total Model Similarity (TMS)

*Total Model Similarity (TMS)* measures the degree of similarity between two models. It is a composite metric based on various individual similarity metrics, as shown in Fig. 2.



**Fig. 2.** Total Model Similarity (TMS)

TMS is computed as the weighted sum of individual similarity measures as shown in Eq. 1.

$$TMS(m_1, m_2) = \frac{\sum_{i=1}^n s_i(m_1, m_2) \cdot w_i}{\sum_{i=1}^n w_i} \quad (1)$$

where

- $m_1, m_2$  : models to be compared,
- $s_i$  : individual similarity measure, i.e.,  $SyMS$ ,  $SeMS$ ,  $MES$ ,  $StMS$ ,
- $w_i$  : weight of individual similarity measure as a numerical value  $> 0$ ,
- $n$  : number of similarity measures, here 4 ( $SyMS$ ,  $SeMS$ ,  $MES$ ,  $StMS$ ).

For  $TMS$  and all similarity metrics specified in this section, a value *threshold* is defined with a statically assigned value, e.g., 0.5. Each similarity measure  $s < threshold$  is, per definition, set to 0. For reasons of readability, the threshold handling is not expressed in every equation explicitly.

The weights of individual similarity measures can be statically assigned, e.g.,  $w_{SyMS}=0.9$ ;  $w_{SeMS}=1.1$ ;  $w_{MES}=3$ ;  $w_{StMS}=1$ . It is obvious that  $0 \leq TMS(m_1, m_2) \leq 1$  if  $0 \leq s_i(m_1, m_2) \leq 1$  for all similarity metrics  $s_i$ .

### 3.2 Syntactic Similarity Metrics (SyMS, SyES)

*Syntactic Model Similarity (SyMS)* and *Syntactic Element Similarity (SyES)* compare the types of modeling languages and language elements based on their meta-models. For example, an EPC is more similar to a UML activity model than to a UML class model. This is because EPC and UML activity models are both expressed in business process modeling languages, whereas UML class models are expressed in a structure modeling language. Analogously, on the element level, an EPC function is more similar to a UML activity than to a UML class. We have implemented a simple similarity metric, namely *Static Similarity Measures*. Similarity measures between model types and model element types are statically assigned, e.g.,  $SyMS(EPC, UML\ activity\ diagram)=0.8$  and  $SyES(EPC\ function, UML\ activity)=0.8$ . Currently we are experimenting with a more complex and flexible similarity metrics, which we call *Meta-Model Reasoning*. In this metric, syntactic similarity measures are determined according to distance measures of model element types in the meta-model.

### 3.3 Semantic Similarity Metrics (SeMS, SeEs)

*Semantic Model Similarity (SeMS)* and *Semantic Element Similarity (SeEs)* compare the meaning of models and model elements, based on their labels and comments. It is a common characteristic of labels in business process models to consist of multiple terms each carrying a semantic meaning but together not forming a grammatically complete sentence. For example, a model named “Travel reservation” is more similar to a model named “Travel booking” than to a model named “Purchasing”. This is because “reservation” is a synonym of “booking”. Analogously, a model element named “request airline reservation” is more similar to an element named “flight reservation” than to an element named “hotel reservation” since “airline” and “flight” are related terms. We have implemented various versions of semantic similarity metrics with increasing complexity and accuracy. Thereby, we are presently concentrating on models with labels in the same natural language.

#### String-Based Matching

In the simplest form, the number of identical terms in the model labels (for *SeMS*), respectively element labels (for *SeEs*) is set in relation to the total number of terms, e.g.,  $SeMS(“Travel\ Reservation”, “Travel\ Booking”)=0.5$ . More sophisticated string-based similarity measures compare characters and their positions in words (e.g., Jaro metric [19]) or the similarity of words in expressions (e.g., Jaccard Metric [20]) [21].

#### Synonym Matching

More complex, but also more accurate metrics include domain-specific knowledge by incorporating a general-language thesaurus like WordNet [22] extendable with further domain-specific thesauri as needed. By rating synonym terms with a pre-defined measure, *SeMS* would return, for example,  $SeMS(“Reservation”, “Booking”)=0.9$ .

### Language-Aware Semantic Matching

The most sophisticated semantic similarity matching procedure we applied is based on a heuristic approach as described in [23]. It combines exact matching, synonym matching and string-based proximity matching for multi-term phrases. Thereby, different ontology matching and linguistic methods are used. For aggregating the results, parameters can be set for weighting the strength of synonyms, stop-words matched and the proximity. In addition, for computing the aggregated result, the term order can be considered, e.g.  $SeMS("Travel Reservation", "Travel Booking")=0.95$ .

### 3.4 Structural Model Similarity (StMS)

The way nodes of a model are connected by arcs, e.g., as loops, can be described as *model structure*. For example, a business process model with three loops is structurally more similar to one with four loops than one with none. We show exemplary one metric for *Structural Model Similarity (StMS)* that measures the difference in the number of loops between the models normalized to the total number of loops, as shown in Eq. 2.

$$StMS(m_1, m_2) = \begin{cases} 1, & l(m_1) = l(m_2) = 0 \\ 1 - \frac{|l(m_1) - l(m_2)|}{l(m_1) + l(m_2)}, & otherwise \end{cases} \quad (2)$$

where

- $m_1, m_2$ : models to be compared,
- $l(m)$ : number of loops in model  $m$ .

It is obvious that  $StMS(m_1, m_2) = 1$  for identical models and  $0 \leq StMS(m_1, m_2) \leq 1$  for all models  $m_1, m_2$ .

### 3.5 Total Element Similarity (TES) and Model Elements Similarity (MES)

*Total Element Similarity (TES)* is measured as the weighted sum of measures resulting from individual similarity metrics. Since this is analogous to Total Model Similarity (TMS) as described in Sect 3.1, we do not show the formula here.

*Model Elements Similarity (MES)* measures how similar all elements of two models are. It sums up the *TES* measures of all element pairs  $e_i, e_j$ , i.e., the Cartesian product  $m_1 \times m_2$ , normalized to the mean number of elements, taking into account a trim factor. See Eq. 3.

$$MES(m_1, m_2) = \min \left( \frac{\sum_{e_i \in m_1, e_j \in m_2} TES(e_i, e_j)}{t_{MES} \cdot (|m_1| + |m_2|)/2}, 1 \right) \quad (3)$$

where

- $m_1, m_2$ : models to be compared,
- $e_i \in m$ : labeled elements in model  $m$ ,
- $|m|$ : number of labeled elements in model  $m$ ,
- $t_{MES}$ : trim factor for MES;  $0 < t_{MES} \leq 1$ .



We only consider labeled elements to allow for their semantic comparison. The trim factor  $t_{MES}$  may statically be assigned as a constant. It allows aligning  $MES$  to the constraints of the model analysis. For example, if 25% correspondence between model elements is regarded sufficient then  $t_{MES}$  should be set to 0.25.  $MES(m_1, m_2) = 1$  if every element  $e_i \in m_1$  has exactly one element  $e_j \in m_2$  with  $TES(e_i, e_j) = 1$ , assuming  $t_{MES} = 1$ . The normalized sum could be greater than 1 if, additionally, elements  $e_i \in m_1$  have other elements  $e'_j \in m_2$  with  $ES(e_i, e'_j) > 0$  and / or  $t_{MES} < 1$ . The use of the minimum function ensures that, in those cases,  $MES$  is bound to 1. It is obvious that  $MES(m_1, m_2) \geq 0$  for all models  $m_1, m_2$  if for all  $e_i, e_j$ ,  $TES(e_i, e_j) \geq 0$ .

## 4 Evaluation

We have implemented the similarity metrics for business process models by using Allegro Common Lisp, AllegroGraph, and AllegroProlog by Franz Inc. and the Semantic Web concept framework as described in [24]. The semantic similarity metrics for model names and element names were imported from the LaSMat system as described in [23]. We have conducted a first empirical validation of the approach presented and its implementation. The results indicate the feasibility of our method and are used as the initial base for proceeding. To avoid biased business process modeling, we have chosen six models from literature, as shown in Table 1.

**Table 1.** Business process models for evaluation

#	Id	Name	Type	Ref.
1.	TRVR	travel reservation	EPC	[13]
2.	TRVB	Travel_Booking	UML Activity Model	[14]
3.	SALS	sales process	EPC	[25]
4.	STOR	standard order handling	EPC	[26]
5.	PROP	procurement-process	EPC	[27]
6.	PROC	procurementprocess	EPC	[27]

The models contain a total of 204 nodes and 229 edges and cover different business domains, namely travel, sales, and procurement. The expected matching pairs are shown in Table 2.

**Table 2.** Expected matching pairs of business process models

Domain	Similar Models
Travel	(1.) TRVR - (2.) TRVB
Sales	(3.) SALS - (4.) STOR
Procurement	(5.) PROP - (6.) PROC

For determining the expected matching pairs, independent domain experts have been chosen to rate the similarity between the business process models, taking into account their domain knowledge.

For the evaluation, the similarity parameters were set as shown in Table 3.

**Table 3.** Similarity parameters for business process model evaluation

Formula	Constants
TMS	$w_{SyMS}=0.9; w_{SeMS}=1.1; w_{MES}=3; w_{SiMS}=1$
TES	$w_{SyEs}=0.9; w_{SeEs}=1.1$
MES	$t_{MES}=0.25$
General	$threshold=0.5$

The rationale for the parameters settings is as follows. Semantic similarity ( $w_{SyMS}=w_{SeEs}=1.1$ ) is rated higher than syntactic similarity ( $w_{SyMS}=w_{SyEs}=0.9$ ). This, and a *threshold* value of 0.5 leads to the desired effect that elements that are syntactically identical (e.g., two EPC functions,  $SyES=1$ ) but semantically non-similar (e.g., “cancel bookings” and “notify customer”,  $SeES=0$ ) are rated as being non-similar ( $TES=0$ ). Total element similarity ( $w_{MES}=3$ ) is rated equally to all similarities on the model level ( $w_{SyMS}+w_{SeMS}+w_{SiMS}=3$ ). 25% of matching elements is considered enough for model elements similarity ( $t_{MES}=0.25$ ). The result of the systematic measurement of TMS for all pairs of process models is shown in Table 4.

**Table 4.** Results for  $TMS(m1, m2)$

$TMS(m1, m2)$	TRVR	TRVB	SALS	STOR	PROP	PROC
TRVR	<b>1.00</b>	<b>0.84</b>	0	0	0	0
TRVB	<b>0.84</b>	<b>1.00</b>	<i>0.79</i>	0	0	0
SALS	0	<i>0.79</i>	<b>1.00</b>	<b>0.55</b>	0	0
STOR	0	0	<b>0.55</b>	<b>1.00</b>	0	0
PROP	0	0	0	0	<b>1.00</b>	<b>0.83</b>
PROC	0	0	0	0	<b>0.83</b>	<b>1.00</b>

Expected similarities (hits) are typeset in bold, non-expected similarities (false positives) are typeset in italics. For assessing the results, we use *Precision* ( $P$ ), *Recall* ( $R$ ) and *F-Measure* ( $F$ ), which are commonly applied measures from Information Retrieval.  $P$  describes the correctness,  $R$  describes the completeness and  $F$  is their weighted harmonic mean. The resulting measures are  $P=0.86$ ;  $R=1.0$ ;  $F=0.92$ . These are satisfying results.

The false positive measure  $TMS(TRVB, SALS)=0.79$  is due to a high  $MES$  measure. When analyzing the individual  $TES$  measures, one finds values like  $TES$ (“*Article reserved*”, “*hotel reservation*”)=0.5. This is, in fact, correct (true positive), since a travel booking is a kind of sales activity whereby a hotel accommodation is being sold.

On the other hand, the value  $TES("Posting", "send invoice")=0.75$  is incorrect (false positive) since "Posting" refers to posting of sold goods and not to posting the invoice.

## 5 Related Work

Reasoning over business process models is an active field of research. In [28], the use of rules is explored for supporting process design and for reasoning about process alternatives when redesigning processes. In [29, 30], ontologies are used for querying and reasoning over business process models in order to support process redesign. The key criterion is the determination of process similarity. In contrast to our approach, a prior developed description of the organization and its processes and process components is required as background information. Aspects of the potentially heterogeneously used domain language used and inter-model relations are not explored. Due to the aim of these research works, metrics for determining the degree of similarity between models have not been addressed.

Some efforts in determining the similarity of process models are concentrating on the aspect of the model languages' semantics for migrating or transforming from one modeling language to another [31–33]. Others are focusing on matching through meta models [34]. These approaches offer methods for analyzing models with respect to their modeling languages, thus leaving out the question of heterogeneously used business language for labeling the model elements.

As a solution for preventing semantic differences concerning the domain language, using agreed-upon sets of terms for creating labels for model elements has been suggested [6, 35]. Such a set may be a business or domain model on the conceptual [35] or business level [36], a domain specific language [5, 37], a domain ontology [38–40], or an enterprise ontology [41, 42]. These approaches assume their prior existence before modeling or a separate top-down development of such domain models. Present suggestions at comparing process models follow this notion by assuming the existence of a separately top-down developed domain ontology for providing mutual understanding of the element's meaning [4, 43]. Works in the field of semantic business process management via model matching over the domain language often rely on such a pre-defined business terminology [12, 44, 45]. However, since the focus is on the alignment of models, metrics for assessing the degree of similarity are not developed in the scope of these works.

Metrics for measuring the similarity of business process models have been developed, concentrating on business process models of a certain type, e.g. EPC [46] or BPMN [47]. Research in the field of consistency regarding model syntax using meta model rules have been done, for example, for determining the soundness of process models [15]. Analysis for semantic consistency concerning the domain language presently focuses on the validation of consistency in the area of Business/IT-Alignment for analyzing appropriate transformation of business process models into executable models [48–50]. In this, differing use of the domain language among process models is not explored.

To our knowledge, the development of measures for process model similarity regardless of the modeling language chosen has not yet been explored. Furthermore, so far, no propositions have been made regarding metrics for determining the

similarity of process models considering both the modeling language and domain language. In this, our approach of measuring semantic similarity as shown could complement the existing efforts in model analysis.

## 6 Conclusion and Outlook

In this paper, we have presented a novel approach for measuring similarity of business process models. The similarity metric encompasses syntactic similarity, structural similarity, and semantic similarity. Different business process modeling languages like EPC or UML Activity modeling are supported. Semantic similarity takes into account the business domain language. We have implemented the approach and applied it to business process models from literature. Our first evaluation results have shown that the method developed allows for computing similarity close to experts' opinion. The benefit lies in leveraging automated computing power in supporting users in determining semantic similarity between arbitrary models.

Future work will include further refinement of our method, especially improving the text mining mechanisms for semantic similarity, allowing for more natural languages than English, analyzing element comments in addition to labels, adding more sophisticated structural similarity metrics, and handling hierarchical business process models. As a long-term goal, our research will focus on applying our methods onto issues surrounding conformance checking of business process models, i.e., measuring the conformance of company-specific business process models to standard reference models.

## References

1. Frank, U.: Informationstechnologie und Organisation. In: Schreyögg, G., Werder, A.v. (eds.) *Handwörterbuch Unternehmensführung und Organisation (HWO)*, pp. 472–480. Schäffer-Poeschel (2004)
2. Scheer, A.-W., Nüttgens, M.: ARIS Architecture and Reference Models for Business Process Management. In: van der Aalst, W.M.P., Desel, J., Oberweis, A. (eds.) *Business Process Management. LNCS*, vol. 1806, pp. 376–389. Springer, Heidelberg (2000)
3. Schmelzer, H.J., Sesselmann, W.: *Geschäftsprozessmanagement in der Praxis*. Hanser, München (2008)
4. Weske, M.: *Business Process Management*. Springer, Heidelberg (2007)
5. Pfeiffer, D.: Constructing Comparable Conceptual Models with Domain Specific Languages. In: Österle, H., Schelp, J., Winter, R. (eds.) *15th Europ. Conf. on Information Systems, ECIS 2007* (2007)
6. Thomas, O., Fellmann, M.: Semantic Business Process Management: Ontology-Based Process Modeling Using Event-Driven Process Chains. *IBIS* 2, 29–44 (2007)
7. Fengel, J.: Semantic Interoperability Enablement in E-Business Modeling. In: Kajan, E. (ed.) *Electronic Business Interoperability*, pp. 331–361. Business Science Reference, Hershey (2011)
8. Hadar, I., Soffer, P.: Variations in Conceptual Modeling: Classification and Ontological Analysis. *Journal of the AIS* 7, 568–592 (2006)
9. Simon, C., Mendling, J.: Integration of Conceptual Process Models by the Example of Event-driven Process Chains. In: Oberweis, A., Weinhardt, C., Gimpel, H. (eds.) *Wirtschaftsinformatik (WI 2007)*. Univ.-Verl. Karlsruhe, Karlsruhe (2007)

10. Frank, U.: Semantische Technologien. *Wirtschaftsinformatik* 52, 49–52 (2010)
11. Lin, Y.: Semantic Annotation for Process Models. Diss. Trondheim, Norway (2008)
12. Hepp, M., Leymann, F., Domingue, J., Wahler, A., Fensel, D.: Semantic Business Process Management. In: Lau, F.C.M., Lei, H., Meng, X., Wang, M. (eds.) *Proc. of ICEBE 2005*, pp. 535–540. IEEE Computer Society, Los Alamitos (2005)
13. Havey, M.: *Essential business process modeling*. O'Reilly, Beijing (2005)
14. IBM: *Business Process Management Samples*, <http://publib.boulder.ibm.com/bpcsamp/>
15. Mendling, J.: *Metrics for Process Models*. Springer, Berlin (2009)
16. Harmon, P.: *Business process change*. Morgan Kaufmann, San Francisco (2003)
17. Scheer, A.W.: *ARIS - Business Process Modeling*. Springer, Berlin (2000)
18. Richter, M.M.: Classification and learning of similarity measures. In: Opitz, O., Lausen, B., Klar, R. (eds.) *Proc. of Information and Classification*. Springer, Heidelberg (1993)
19. Jaro, M.A.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa. *Journal of the American Statistical Association* 84, 351–357 (1989)
20. Jaccard, P.: The Distribution of the Flora in the Alpine Zone. *The New Phytologist* 11, 37–50 (1912)
21. Cohen, W., Ravikumar, P., Fienberg, S.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: *Proc. of IJCAI 2003 Workshop on Information Integration on the Web (IIWeb 2003)*, pp. 73–78 (2003)
22. Princeton University: *WordNet - About WordNet*, <http://wordnet.princeton.edu/>
23. Fengel, J., Reinking, K.: Sprachbezogener Abgleich der Fachsemantik in heterogenen Geschäftsprozessmodellen. In: *Proc. of Fachtagung Modellierung 2012*, March 14–16 (to appear, 2012)
24. Humm, B., Korobov, A.: Introducing Layers of Abstraction to Semantic Web Programming. In: Meersman, R., Dillon, T., Herrero, P. (eds.) *OTM-WS 2011. LNCS*, vol. 7046, pp. 412–423. Springer, Heidelberg (2011)
25. Scheer, A.-W.: *Business Process Engineering*, 2nd edn. Springer, Heidelberg (1997)
26. Curran, T.A., Ladd, A.: *SAP R/3 Business Blueprint: Understanding Enterprise Supply Chain Management*. 2nd ed. Pearson Education (1998)
27. Simon, C.: *Negotiation Processes: The Semantic Process Language and Applications*. Shaker, Aachen (2008)
28. Yu, E.S., Mylopoulos, J.: Using Goals, Rules and Methods to Support Reasoning in Business Process Reengineering. *Intelligent Systems in Accounting, Finance & Management* 5 (1996)
29. Markovic, I.: Advanced Querying and Reasoning on Business Process Models. In: Abramowicz, W., Fensel, D. (eds.) *BIS 2008. LNBIP*, vol. 7, pp. 189–200. Springer, Heidelberg (2008)
30. Missikoff, M., Proietti, M., Smith, F.: Reasoning on Business Processes and Ontologies in a Logic Programming Environment. In: Missikoff, M., Velardi, P. (eds.) *Proc. of 3rd Interop-Vlab. Workshop on Enterprise Interoperability*, vol. 653, pp. 17–21. CEUR (2010)
31. Murzek, M., Kramler, G.: The Model Morphing Approach – Horizontal Transformations between Business Process Models. In: Nummenmaa, J. (ed.) *Proc. of the 6th Intern. Conf. on Perspectives in Business Information Research - BIR 2007*, pp. 88–103 (2007)
32. Gehlert, A.: *Migration fachkonzeptueller Modelle*. Logos-Verl., Berlin (2007)
33. Kenschke, D., Quix, C., Chatti, M.A., Jarke, M.: *GeRoMe: A Generic Role Based Metamodel for Model Management*. In: Spaccapietra, S., Atzeni, P., Fages, F., Hacid, M.-S., Kifer, M., Mylopoulos, J., Pernici, B., Shvaiko, P., Trujillo, J., Zaihrayeu, I., et al. (eds.) *Journal on Data Semantics VIII. LNCS*, vol. 4380, pp. 82–117. Springer, Heidelberg (2007)

34. Kappel, G., Kramler, G., Kapsammer, E., Reiter, T., Retschitzegger, W., Schwinger, W.: ModelCVS - A Semantic Infrastructure for Model-based Tool Integration. Wien (2005)
35. Kugeler, M., Rosemann, M.: Fachbegriffsmodellierung für betriebliche Informationssysteme und zur Unterstützung der Unternehmenskommunikation. *Informationssystem Architekturen* 5, 8–15 (1998)
36. OMG: MDA Guide Version 1.0.1, <http://www.omg.org/docs/omg/03-06-01.pdf>
37. Kurtev, I., Bézivin, J., Jouault, F., Valduriez, P.: Model-based DSL Frameworks. In: *OOPSLA 2006: Companion to the 21st ACM SIGPLAN Symposium on Object-Oriented Programming Systems, Languages, and Applications*, pp. 602–616 (2006)
38. Hepp, M., Roman, D.: An Ontology Framework for Semantic Business Management. In: Oberweis, A., Weinhardt, C., Gimpel, H. (eds.) *Wirtschaftsinformatik (WI 2007)*. Univ.-Verl. Karlsruhe, Karlsruhe (2007)
39. Saeki, M., Kaiya, H.: On Relationships among Models, Meta Models and Ontologies. In: Gray, J., Tolvanen, J.P., Sprinkle, J. (eds.) *Proc. of the 6th OOPSLA Workshop on Domain-Specific Modeling* (2006)
40. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: *Ontological Engineering*. Springer, London (2004)
41. Grüninger, M., Atefi, K., Fox, M.A.: Ontologies to Support Process Integration in Enterprise Engineering. *Computational & Mathematical Organization Theory* 6, 381–394 (2000)
42. Dietz, J.L.G., Hoogervorst, J.A.P.: Enterprise Ontology in Enterprise Engineering. In: Wainwright, R.L., Haddad, H.M. (eds.) *Proc. of the 2008 ACM Symposium on Applied Computing*, pp. 572–579. ACM, New York (2008)
43. Becker, J., Pfeiffer, D.: Solving the Conflicts of Distributed Process Modelling – Towards an Integrated Approach. In: Golden, W., Acton, T., Conboy, K., van der Heijden, H., Tuunainen, V.K. (eds.) *16th Europ. Conf. on Inform. Systems (ECIS 2008)*, pp. 1555–1568 (2008)
44. Brockmans, S., Ehrig, M., Koschmider, A., Oberweis, A., Studer, R.: Semantic Alignment of Business Processes. In: Manolopoulos, Y., Filipe, J., Constantopoulos, P., Cordeiro, J. (eds.) *Proc. of the 8th Intern. Conf. on Enterprise Information Systems (ICEIS 2006)*, pp. 197–203. INSTICC, Setúbal (2006)
45. Thomas, O., Fellmann, M.: Semantische Prozessmodellierung – Konzeption und informationstechnische Unterstützung einer ontologiebasierten Repräsentation von Geschäftsprozessen. *Wirtschaftsinformatik* 51, 506–518 (2009)
46. Dijkman, R., van Dongen, B., Dumas, M., Käärik, R., Mendling, J.: Similarity of Business Process Models: Metrics and Evaluation. Working Paper (2009)
47. Kunze, M., Weske, M.: Metric Trees for Efficient Similarity Search in Large Process Model Repositories. In: zur Muehlen, M., Su, J. (eds.) *BPM 2010 Workshops. LNBIP*, vol. 66, pp. 535–546. Springer, Heidelberg (2011)
48. Born, M., Hoffmann, J., Kaczmarek, T., Kowalkiewicz, M., Markovic, I., Scicluna, J., Weber, I., Zhou, X.: Supporting Execution-Level Business Process Modeling with Semantic Technologies. In: Zhou, X., Yokota, H., Deng, K., Liu, Q. (eds.) *DASFAA 2009. LNCS*, vol. 5463, pp. 759–763. Springer, Heidelberg (2009)
49. Weber, I., Hoffmann, J., Mendling, J.: Beyond Soundness: On the Semantic Consistency of Executable Process Models. In: Pahl, C. (ed.) *IEEE Sixth Europ. Conf. on Web Services 2008*, pp. 102–111. IEEE, Piscataway (2008)
50. Morrison, E.D., Menzies, A., Koliadis, G., Ghose, A.K.: *Business Process Integration: Method and Analysis*. Technical Report (2010)

# Data- and Resource-Aware Conformance Checking of Business Processes

Massimiliano de Leoni, Wil M.P. van der Aalst, and Boudewijn F. van Dongen

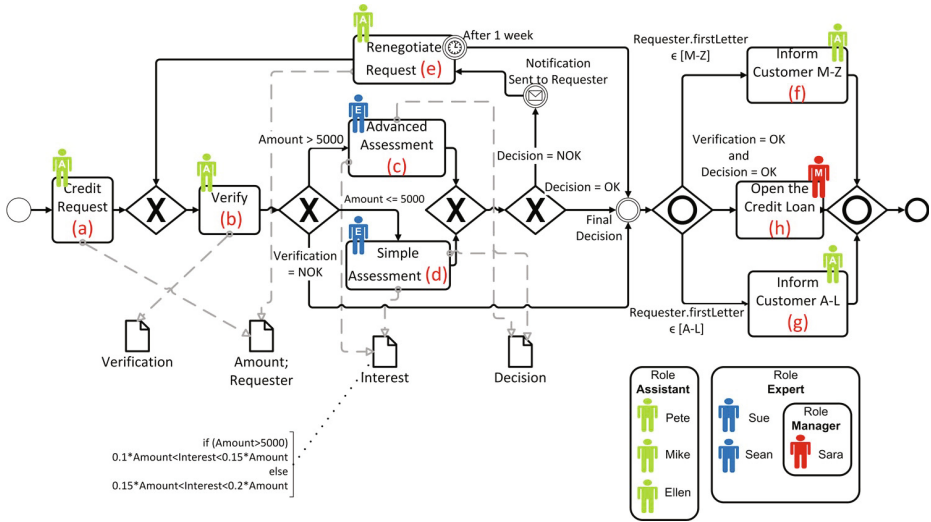
Eindhoven University of Technology, Eindhoven, The Netherlands  
{m.d.leoni,w.m.p.v.d.aalst,b.f.v.dongen}@tue.nl

**Abstract.** Process mining is not restricted to process discovery and also includes *conformance checking*, i.e., checking whether observed behavior recorded in the event log matches modeled behavior. Many organizations have descriptive or normative models that do not adequately describe the actual processes. Therefore, a variety of techniques for conformance checking have been proposed. However, *all of these techniques focus on the control-flow and abstract from data and resources*. This paper describes an approach that aligns event log and model while taking all perspectives into account (i.e., also data and resources). This way it is possible to quantify conformance and analyze differences between model and reality. The approach has been implemented in ProM and evaluated using a variety of model-log combinations.

## 1 Introduction

Modern organizations are centered around the processes needed to deliver products and services in an efficient and effective manner. Organizations that operate at a higher process maturity level use formal/semiformal models (e.g., UML, EPC, BPMN and YAWL models) to document their processes. In some case these models are used to configure process-aware information systems (e.g., WFM or BPM systems). However, in most organizations process models are not used to enforce a particular way of working. Instead, process models are used for discussion, performance analysis (e.g., simulation), certification, process improvement, etc. However, reality may deviate from such models. People tend to focus on idealized process models that have little to do with reality. This illustrates the importance of *conformance checking* [1][2][3].

An important enabler for conformance checking is the availability of event data in modern organizations. Even though processes are typically not enforced by a process-aware information system, still most events are recorded. Consider for example a hospital. Medical doctors are not controlled by some BPM system. However, many events are recorded, e.g., blood tests, X-ray images, administered drugs, surgery, etc. all result in events that can be linked to a particular patient. Digital data is everywhere – in every sector, in every economy, in every organization, and in every home – and will continue to grow exponentially. MGI estimates that enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks [4]. The growing availability of event data is an important enabler for conformance checking.



**Fig. 1.** BPMN diagram of a data and resource-aware process to manage credit requests to buy home appliances. In the remainder, data objects are simply referred with the upper-case initials, e.g.,  $V=Verification$ , and activity names by the letter in brackets, e.g.  $a=Credit Request$ .

Along with *process discovery* (learning process models from logs) and *process enhancement* (e.g., extending process models with bottleneck information based on timestamps in event logs), conformance checking belongs to the area of *Process Mining* [5], which is a relatively young research discipline that sits between computational intelligence and data mining on the one hand, and process modeling and analysis on the other hand.

Conformance checking techniques take an event log and a process model and compare the observed behavior with the modeled behavior. There are different dimensions for comparing process models and event logs. In this paper, we focus on the *fitness* dimension: a model with good fitness allows for most of the behavior seen in the event log. A model has a *perfect* fitness if all traces in the log can be replayed by the model from beginning to end. Other quality dimensions are *simplicity*, *precision*, and *generalization* [11].

Various conformance checking techniques have been proposed in recent years [1, 2, 3, 6, 7, 8, 9, 10, 11]. However, all of the techniques described in literature focus on the control flow, i.e. the ordering of activities. They do not take into account other perspectives, such as resources and data. For example, when an activity is executed by the wrong person it is important to detect such a deviation. Conformance checking techniques need to detect that an activity reserved for gold customers is executed for silver customers. Note that information about cases and data is readily available in today's event logs. The routing of a case may depend on data, i.e., a check needs to be performed for claims over 5000. Therefore, it is important to look at the combination of all perspectives.

In a process model each case, i.e. a process instance, is characterized by its case attributes. Paths taken during the execution may be governed by guards and conditions



defined over such attributes. Process models also define, for each attribute, its domain, i.e. the values that can be given. Moreover, process models prescribe which attributes every activity can read or write. Last but not least, process models also describe which resources are allowed to execute which activities. An activity is typically associated with a particular role, i.e., a selected group of resources. Moreover, there may be additional rules such as the “four-eyes principle” which does not allow for the situation where the same resource executes two related tasks. If the data and the resource perspective are not considered, process executions can apparently be fully conforming, whereas actually they are not. Let us consider the following example:

**Example 1.** *A credit institute has activated a process to deal with loans requested by clients. These loans can be used to buy small home appliances (e.g., fridges, TVs, high-quality digital sound systems). A customer can apply for a loan through a shop clerk. The clerk prepares the request by filling out the form and attaching documents that prove the capability to pay off the loan. Upon receiving a new request, the credit institute opens a new case of the process in Figure 7. Dotted lines going from activities to data objects indicate the data objects (i.e., the attributes) that activities are allowed to change the value of. The resource perspective is specified by defining the role that participants need to have in order to execute the corresponding activity.*

Let us also consider the following trace where attribute  $E$  stands for *Executor* and denotes the activity executor<sup>1</sup>

$\langle\langle \mathbf{a}, \{A = 3000, R = \text{Michael}, E = \text{Pete}\} \rangle\rangle, \langle\langle \mathbf{b}, \{V = \text{OK}, E = \text{Sue}, A = 3000, R = \text{Michael}\} \rangle\rangle,$   
 $\langle\langle \mathbf{c}, \{I = 530, D = \text{OK}, E = \text{Sue}, A = 3000, R = \text{Michael}\} \rangle\rangle, \langle\langle \mathbf{f}, \{E = \text{Pete}, A = 3000, R = \text{Michael}\} \rangle\rangle.$

Existing conformance checking techniques [2,3,6,7,8,9,10,11] would only consider the control flow and ignore the decision points, values assigned to attributes and resources. Hence, the given trace would be considered as perfectly fitting. The approach proposed in this paper also considers the data and resource perspectives. For example, using our techniques, we can discover violations of rules, such as: (i) activity **c** should not be executed since the loan amount is not greater than 5000 (conversely, activity **d** should); (ii) for the considered credit loan, the interest should not be more 450 Euros and, hence, proposing an interest of 530 Euros is against the credit-institute’s policy for small loans; (iii) ‘Sue’ is not authorized to execute activity **b** since she cannot play role *Assistant*; (iv) activity **h** has not been executed and, hence, the decision cannot be positive. The approach we propose is based on the principle of finding an *alignment* of event log and process model. The events in the traces are mapped to the execution of activities in the process model. Such an alignment shows how the event log can be *replayed* on the process model. In [12] an alignment-based conformance checking techniques is described. However, this approach is limited to the control-flow perspective. This paper extends [12] by also taking the data and resource perspectives into account.

We allow costs to be assigned to every potential deviation. Some deviations are more severe than others and the severity can also be influenced by the point in the process when these occur, e.g., skipping a notification activity is more severe for gold customers. Our approach uses the A\* algorithm [13] to find, for each trace in the event log,

<sup>1</sup> Notation  $(act, \{attr_1 = val_1, \dots, attr_n = val_n\})$  is used to denote the occurrence of activity  $act$  in which attributes  $attr_1, \dots, attr_n$  are assigned values  $val_1, \dots, val_n$ , respectively.

the process execution, among those possible, whose deviations from the log trace has the lowest overall cost. In order to keep the technique as general as possible, we have developed it as independent of both the actual language in which business processes are described and the log format. Together with measuring the degree of conformance, the technique highlights where deviations occur thereby showing the control-flow, data and resource perspectives. In particular, among the different types of deviations that the technique can diagnose, it is capable to compute how much a value assignment to an attribute deviates. Similarly, from the resource viewpoint, the techniques pinpoints which resources and activities more often violate the authorization.

Section 2 illustrates a formalism that abstracts from the actual log and process notation and focuses on the behavior described by the model and recorded in the event logs. Section 3 shows how constructing an optimal alignment of process model and event log can be used to diagnose non-conformance and quantify the fitness. Section 4 elaborates the adaptation of the A\* algorithm to solve the problem of conformance checking. Section 5 describes our implementation of this new approach in ProM. Moreover, experimental results are given. Finally, Section 6 concludes the paper, describing future directions of improvement.

## 2 The General Framework

Typically, any process model, such as the BPMN diagram in Figure 1, relies on constructs such as parallel split nodes, synchronization nodes, decision/choice nodes, conditions, merge nodes, etc. However, the model description can be “expanded” into a (possible infinite) set of (potentially arbitrarily long) traces yielding to a final state, i.e. the set of admissible behaviors. Each trace can be seen as a sequence of *execution steps*, each of which corresponds to the execution of a given process activity. Usually, a process model also defines a set of attributes together with their domain (i.e., the values that can be given). An activity is allowed to read and write attributes in a predefined manner.

Let  $A, V$  be, respectively, the finite set of activities and attributes. For all attributes  $v \in V$ , let us denote with  $\text{domAttr}(v)$  the set of values allowed for  $v$  (i.e., the attribute domain). Let be  $U = \bigcup_{v \in V} \text{domAttr}(v)$ . A **execution step**  $s = (a_s, \varphi_s)$  consists of an executed activity  $a_s$  and a function that denotes an assignment of values to process attributes:  $\varphi_s \in V \rightarrow U$  s.t.  $\forall v \in \text{dom}(\varphi_s). \varphi_s(v) \in \text{domAttr}(v)$ .<sup>2</sup> Let  $S$  be the set of possible execution steps. A **process**  $\mathcal{P}$  is the set of all admissible execution traces:  $\mathcal{P} \subseteq S^*$ . For each execution step  $s = (a_s, \varphi_s)$ , we use function  $\#_{act}(s) = a_s$  to extract the activity associated to the execution step.

*Resources are taken into account by “reserving” a special attribute to carry the executor information.* Each value assignment to attributes can either be a read or write operation and the semantics depends on the executed activity and event log. For instance, let us consider the trace in Section 1 and the first two execution steps  $s' = (\mathbf{a}, \{A = 3000, R = \text{Michael}, E = \text{Pete}\})$  and  $s'' = (\mathbf{b}, \{V = \text{OK}, E = \text{Sue}, A = 3000, R = \text{Michael}\})$ . The assignment  $A = 3000$  for  $s'$  denotes that the execution of step  $s'$  provokes an assignment of value 3000 to attribute  $A$ . Conversely,  $A = 3000$

<sup>2</sup> The domain of a function  $f$  is denoted by  $\text{dom}(f)$ .

for  $s''$  indicates that, during the execution of step  $s''$ , the value 3000 has been read for attribute  $A$ . In the remainder, we focus on the writing operations. It is obvious to see that our approach can be extended to distinguish between read and write operations.

An event log contains events associated to cases, i.e., process instances. Each case follows a trace of events. Each trace records the execution of a process instance. Different instances may follow the same trace. Therefore, an **event log** is a multi-set of traces, i.e.,  $\mathcal{L} \in \mathbb{B}(S^*)$ <sup>3</sup>

### 3 Aligning Event Log and Process Model

Conformance checking requires an *alignment* of event log  $\mathcal{L}$  and process model  $\mathcal{P}$ : the events in the event log need to be related to model elements and vice versa. Such an alignment shows how the event log can be replayed on the process model. This is far from being trivial since the log may deviate from the model and not all activities may have been modeled and recorded.

We need to relate “moves” in the log to “moves” in the model in order to establish an alignment between a process model and an event log. However, it may be the case that some of the moves in the log cannot be mimicked by the model and vice versa. We explicitly denote “no move” by  $\perp$ . For convenience, we introduce the set  $S_{\perp} = S \cup \{\perp\}$ .

One step in an *alignment* is represented by a pair  $(s', s'') \in (S_{\perp} \times S_{\perp}) \setminus \{(\perp, \perp)\}$  such that

- $(s', s'')$  is a *move in log* if  $s' \in S$  and  $s'' = \perp$ ,
- $(s', s'')$  is a *move in process* if  $s' = \perp$  and  $s'' \in S$ ,
- $(s', s'')$  is a *move in both* if  $s' \in S$  and  $s'' \in S$ .

$S_A = (S_{\perp} \times S_{\perp}) \setminus \{(\perp, \perp)\}$  is the set of all *legal moves* where the first and the second element of every pair denote possible moves in the log and in the process, respectively.

The **alignment** of two execution traces  $\sigma', \sigma'' \in S^*$  is a sequence  $\gamma \in S_A^*$  such that, ignoring all occurrences of  $\perp$ , the projection on the first element yields to  $\sigma'$  and the project on the second yields to  $\sigma''$ . In particular,  $\gamma$  is a **complete alignment** if  $\sigma' \in \mathcal{L}$  and  $\sigma'' \in \mathcal{P}$ .

In order to define the severity of a deviation, we introduce a cost function on legal moves:  $\kappa \in S_A \rightarrow \mathbb{R}_0^+$ . The costs of each legal move depends on the specific model and process domain and, hence, cost function  $\kappa$  needs to be defined ad-hoc for every specific case. The cost function can be generalized to alignments  $\gamma$  as the sum of the cost of each individual move:  $\mathcal{K}(\gamma) = \sum_{(s', s'') \in \gamma} \kappa(s', s'')$ .

**Example 1 (cont.).** *When checking for conformance, the business analysts repute more severe the misconformances on activities that are concerned with interactions with customers, since they can undermine the reputation of the credit institute. Therefore, every alignment step between  $\perp$  and an execution step for activities  $\mathbf{c}$  or  $\mathbf{d}$  is given a cost 1, whereas a cost of 10 is given to alignment steps between  $\perp$  and execution steps for any activity different from  $\mathbf{c}$  and  $\mathbf{d}$ :*

$$\forall s \in S. \bar{\kappa}(s, \perp) = \bar{\kappa}(\perp, s) = \begin{cases} 1 & \text{if } \#_{act}(s) \in \{\mathbf{c}, \mathbf{d}\} \\ 10 & \text{if } \#_{act}(s) \notin \{\mathbf{c}, \mathbf{d}\} \end{cases}$$

<sup>3</sup>  $\mathbb{B}(X)$  the set of all multi-sets over  $X$ .

$\gamma_1$	
a, {A=3000,R=Michael,E=Pete}	a, {A=3000,R=Michael,E=Pete}
b, {V=OK,E=Pete}	b, {V=OK,E=Sue}
c, {I=530,D=OK,E=Sue}	$\perp$
$\perp$	d, {I=599,D=NOK,E=Sue}
f, {E=Pete}	f, {E=Pete}

$\gamma_2$	
a, {A=3000,R=Michael,E=Pete}	a, {A=3000,R=Michael,E=Pete}
b, {V=OK,E=Pete}	$\perp$
$\perp$	b, {V=OK,E=Sue}
c, {I=530,D=OK,E=Sue}	$\perp$
$\perp$	d, {I=599,D=NOK,E=Sue}
f, {E=Pete}	f, {E=Ellen}

$\gamma_3$	
a, {A=3000,R=Michael,E=Pete}	a, {A=3000,R=Michael,E=Pete}
b, {V=OK,E=Pete}	b, {V=OK,E=Sean}
c, {I=530,D=OK,E=Sue}	$\perp$
$\perp$	d, {I=500,D=NOK,E=Sue}
f, {E=Pete}	f, {E=Pete}

$\gamma_4$	
a, {A=3000,R=Michael,E=Pete}	a, {A=5001,R=Michael,E=Pete}
b, {V=OK,E=Pete}	b, {V=OK,E=Sean}
c, {I=530,D=OK,E=Sue}	c, {I=530,D=NOK,E=Sue}
f, {E=Pete}	f, {E=Pete}

**Fig. 2.** Four possible alignments of the log trace described in Section 1 and the process model in Figure 1

Let  $\text{Diff}(s', s'')$  be the set of attributes to which both steps  $s'$  and  $s''$  assign a value, but a different one. Every move in both is assigned a cost as follows:

$$\forall s', s'' \in S. \bar{\kappa}(s', s'') = \begin{cases} 0.2 \cdot \|\text{Diff}(s', s'')\| & \text{if } \#_{\text{act}}(s') = \#_{\text{act}}(s'') \wedge \#_{\text{act}}(s') \in \{c, d\} \\ 3 \cdot \|\text{Diff}(s', s'')\| & \text{if } \#_{\text{act}}(s') = \#_{\text{act}}(s'') \wedge \#_{\text{act}}(s') \notin \{c, d\} \\ \infty & \text{otherwise} \end{cases}$$

The idea is that moves in both with different value assignment to attributes are given a higher cost for activities  $c$  and  $d$  rather than for any other activity. Let us consider again the log trace given in Section 1. Figure 2 shows four possible alignments. It is easy to check that  $\mathcal{K}(\gamma_1) = \mathcal{K}(\gamma_3) = 0+3+1+1+0 = 5$ ,  $\mathcal{K}(\gamma_2) = 0+10+10+1+1+2 = 24$  and  $\mathcal{K}(\gamma_4) = 3+2+0.6+0 = 5.6$  and, hence, alignments  $\gamma_1$  and  $\gamma_3$  are certainly better than  $\gamma_2$  and  $\gamma_4$ .

So far we have considered single complete alignments. However, given a log trace  $\sigma_L \in \mathcal{L}$ , our goal is to find a complete alignment of  $\sigma_L$  and  $\mathcal{P}$  which minimizes the cost with respect to all  $\sigma'_P \in \mathcal{P}$ . We refer to it as an optimal alignment. Let  $\Gamma_{\sigma_L, \mathcal{P}}$  be the set of all complete alignments of  $\sigma_L$  and  $\mathcal{P}$ . The alignment  $\gamma \in \Gamma_{\sigma_L, \mathcal{P}}$  is an **optimal alignment** if  $\forall \gamma' \in \Gamma_{\sigma_L, \mathcal{P}}. \mathcal{K}(\gamma) \leq \mathcal{K}(\gamma')$ . Note that there may exist several optimal alignments, i.e. several complete alignments of the same minimal cost.

**Example 1 (cont.).** For this example, using the cost function  $\bar{\kappa}$  defined above,  $\gamma_1$  and  $\gamma_3$  are both optimal alignments. Of course, the set of optimal alignments depends on the cost function  $\kappa$ . For instance, let us consider a cost function  $\hat{\kappa}$  s.t.  $\forall s \in S. \hat{\kappa}(\perp, s) = \hat{\kappa}(s, \perp) = 10$  and  $\forall s', s'' \in S. \hat{\kappa}(s', s'') = \bar{\kappa}(s', s'')$ . Using  $\hat{\kappa}$  as cost function, the alignment  $\gamma_4$  would be optimal with  $\mathcal{K}(\gamma_4) = 5.6$ , whereas alignments  $\gamma_1$  and  $\gamma_3$  would no more be optimal since  $\mathcal{K}(\gamma_1) = \mathcal{K}(\gamma_3) = 22$ .

In the next section we propose an approach to create an optimal alignment with respect to a custom cost function  $\kappa$ . The approach is based on the A\* algorithm, i.e. an algorithm intended to find the path with the lowest overall cost between two nodes in a direct graph with costs associated to nodes. We have adapted it to derive one of the optimal alignments.

## 4 The A\* Algorithm for Conformance Checking

The A\* algorithm, initially proposed in [13], aims at finding a path in a graph  $V$  from a given *source* node  $v_0$  to any node  $v \in V$  in a target set. With every node  $v$  of graph  $V$  there is an associated cost, which is determined by an *evaluation* function  $f(v) = g(v) + h(v)$ , where

- $g : V \rightarrow \mathbb{R}_0^+$  is a function that returns the smallest path cost from  $v_0$  to  $v$ ;
- $h : V \rightarrow \mathbb{R}_0^+$  is a heuristic function that estimates the path cost from  $v$  to its preferred target node.

Function  $h$  is said to be *admissible* if it returns a value that underestimates the distance of a path from a node  $v'$  to its preferred target node  $v''$ , i.e.  $h(v') \leq g(v'')$ . If  $h$  is admissible, A\* finds a path that is guaranteed to have the overall lowest cost.

The A\* algorithm keeps a priority queue of nodes to be visited: higher priority is given to nodes with lower costs so as to traverse those with the lowest costs at first. The algorithm works iteratively: at each step, the node  $v$  with lowest cost is taken from the priority queue. If  $v$  belongs to the target set, the algorithm ends returning node  $v$ . Otherwise,  $v$  is expanded: every successors  $v'$  is added to priority queue with a cost  $f(v')$ .

We employ A\* to find any of the optimal alignments between a log trace  $\sigma_L \in S^*$  and a Process Model  $\mathcal{P}$ . In order to be able to apply A\*, an opportune search space needs to be defined. Every node  $\gamma$  of the search space  $V$  is associated to a different alignment that is a prefix of some complete alignment of  $\sigma_L$  and  $\mathcal{P}$ . Since a different alignment is also associated to every node and vice versa, later on we use the alignment to refer to the associated state. The source node is empty alignment  $\gamma_0 = \langle \rangle$  and the set of target nodes includes every complete alignment of  $\sigma_L$  and  $\mathcal{P}$ .

Let us denote the length of a sequence  $\sigma$  with  $\|\sigma\|$ . Given a node/alignment  $\gamma \in V$ , the search-space successors of  $\gamma$  include all alignments  $\gamma' \in V$  obtained from  $\gamma$  by concatenating exactly one move step. Let us consider a custom cost function  $\kappa$  and denote with  $\kappa^{\min}$  the smallest value returned by  $\kappa$  that is greater than 0. Given an alignment  $\gamma \in V$  of  $\sigma'_L$  and  $\sigma'_P$ , the cost of path from the initial node to node  $\gamma \in V$  is:

$$g(\gamma) = \kappa^{\min} \cdot \|\sigma'_L\| + \mathcal{K}(\gamma).$$

It is easy to check that, given a log trace  $\sigma_L$  and two complete alignments  $\gamma'_C$  and  $\gamma''_C$  of  $\sigma_L$  and  $\mathcal{P}$ ,  $\mathcal{K}(\gamma'_C) < \mathcal{K}(\gamma''_C)$  iff  $g(\gamma'_C) < g(\gamma''_C)$  and  $\mathcal{K}(\gamma'_C) = \mathcal{K}(\gamma''_C)$  iff  $g(\gamma'_C) = g(\gamma''_C)$ . Therefore, an optimal solution returned by the A\* algorithm coincides with an optimal alignment. Term  $\kappa^{\min} \cdot \|\sigma'_L\|$ , which does not affect the optimality, has been added because it allows us to define a more efficient admissible heuristics. Given an alignment  $\gamma \in V$  of  $\sigma'_L$  and  $\sigma'_P$ , we employ the following heuristics:

$$h(\gamma) = \kappa^{\min} \cdot (\|\sigma_L\| - \|\sigma'_L\|)$$

For alignment  $\gamma$ , the number of steps to add in order to reach a complete alignment is lower bounded by the number of execution steps of trace  $\sigma_L$  that have not been included yet in the alignment, i.e.  $\|\sigma_L\| - \|\sigma'_L\|$ . Since the additional cost to traverse a single node is at least  $\kappa^{\min}$ , the cost to reach a target node is at least  $h(\gamma)$ , corresponding to the case when the part of the log trace that still needs to be included in the alignment fits in full.

## 5 Implementation and Experiments

The *Data-Aware Conformance Checker* is implemented as a software plug-in of ProM, a generic open-source framework for implementing process mining tools in a standard environment [14]. The plug-in takes as input a process model and a log and, by employing the techniques described in Section 4, answers to the conformance-checking questions expressed in the Section 1.

*Extended Casual Nets.* Our data-aware conformance-checking engine is completely independent of the process modeling language. As a proof of concept, we have used *Causal Nets* as concrete language to represent process models and extended it in order to describe the aspects related to the data and resource perspective. While Casual Nets without the data and resource perspective are thoroughly described in [11], space limitations prevent us from giving here a full formalization for their extension with these perspectives. A Casual Net extended with data is a graph where nodes represent activities and arcs represent causal dependencies. Each activity has a set of possible *input bindings* and

*output bindings*. The occurrence of an activity is represented by an *activity binding*  $(a, ab^I, ab^O, \phi)$ , which denotes the occurrence of activity  $a$  with input binding  $ab^I$  and output binding  $ab^O$  and data binding function  $\phi$ , where data attributes have global scopes. The input and output bindings include the activities that precede and succeed the occurrence of activity  $a$ . If there exists an attribute  $v \in \text{dom}(\phi)$  and a value  $u$  such that  $\phi(v) = u$ , the occurrence of  $a$  provokes to overwrite the value of attribute  $u$  with  $u$ . The definition of a process  $\mathcal{P}$  in Section 2 is also applicable to Extended Casual Nets: there exists a distinct process trace for each legal sequence of activity bindings that ends with the final activity.<sup>4</sup> Given a valid sequence of activity bindings, the corresponding process trace contains a different execution step  $(a, \phi)$  for each activity binding  $(a, ab^I, ab^O, \phi)$ . And the order of the execution steps in a process trace complies the order of activity bindings in the corresponding activity-bindings sequences.

**Example 2.** Figure 3 shows an example of a Causal Net extended with data. There is a set of  $n$  different process attributes  $X_1, \dots, X_n$ , each of which is a natural number between 1 and  $m$ . Node  $S$  is the starting activity: it has no input binding and one output binding, which is the set  $\{A_1, \dots, A_n\}$  of activities. This means activity  $S$  is followed by activities  $A_1, \dots, A_n$  executed in any order (i.e., AND-split). Activity  $A_i$  is associated a guard  $X_i \geq 0$ ; when an attribute, e.g.  $X_i$ , is annotated with the prime symbol in a

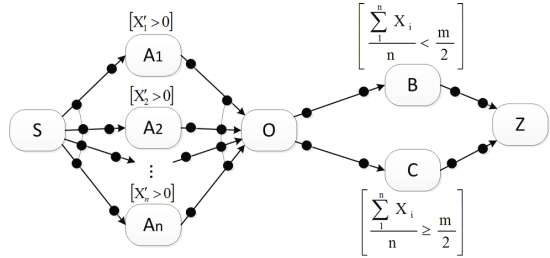
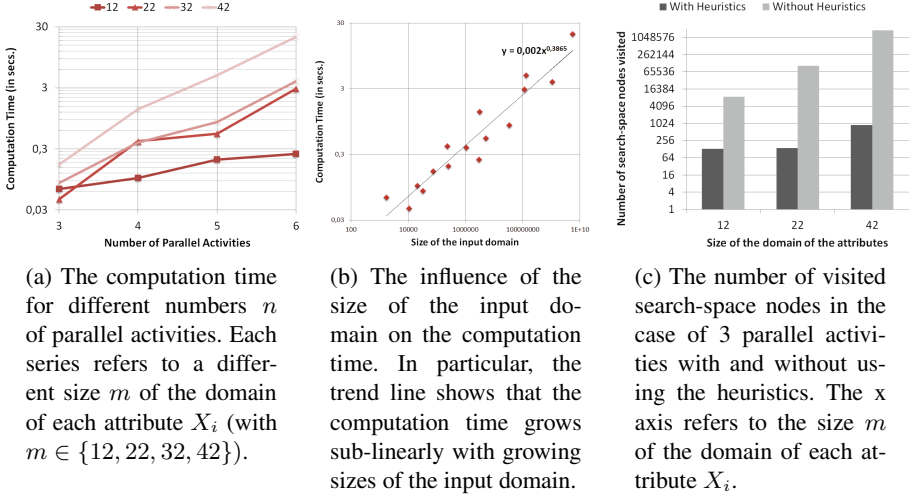


Fig. 3. An Extended Causal Net for Example 2

<sup>4</sup> The lack of space does not allow us to elaborate more the concept of “legal sequence”. In a few words, a sequence of activity bindings is valid if every predecessor activity and successor activity always agree on their bindings as well as the guards always hold in relation with the value assignments given. Interested readers can also refer to [11].



**Fig. 4.** The results of the experiments conducted on Example 2

guard, the activity, e.g.  $A_i$ , is prescribed to update the value of the attribute. And the written value must not violate the guards, e.g.  $X_i$  has to be assigned a non-negative value. Activity  $O$  is characterized by an input binding  $\{A_1, \dots, A_n\}$ , which means that  $O$  can only be executed after all activities  $A_1, \dots, A_n$  have been (i.e., AND-join). Two possible output bindings are modeled for  $O$ :  $B$  and  $C$ . Therefore,  $O$  is followed by either  $B$  or  $C$  (i.e., XOR-split).  $B$  and  $C$  are associated with two guards indicating that activities  $B$  or  $C$  can follow  $O$  if the average of values for  $X_1, \dots, X_n$  is less than  $m/2$  or, vice versa, greater or equal to  $m/2$ .

*Experiments.* As input for experiments, we generated event logs by modeling the Examples 2 in CPN Tools (<http://cpntools.org>) and simulating the model. In particular, we considered all combinations in which the number  $n$  of parallel activities ranges from 3 to 6 and each attribute can be given a value between 1 and  $m$ , with  $m \in \{12, 22, 32, 42\}$ .

For each combinations, the log is composed by 6 different traces, generated in a random fashion but perfectly fitting. In order to make the conformance-checking analysis more challenging, from the generated logs, we have removed every occurrence of activity  $A_1$  as well as we have swapped every occurrence of activity  $B$  and  $C$ . Moreover, we have set the cost of moving only in the process or in the log three times bigger than moving in both with different value assignments to attribute. In this way, complete alignments that contain move only in the process and in the log are always given a cost higher than move in both with different value assignments. Therefore, in order to find an optimal alignment, the conformance checker needs to find different value assignments to attributes  $X_1, \dots, X_n$  from what encountered in the log. In this way, moves only in the log or in the process can be avoided for  $B$  or  $C$ .

Figure 4 illustrates the results of the experiments. The graph in Figure 4a shows the computation time to find optimal alignments for different numbers  $n$  of parallel activities and for different sizes  $m$  of the domain of each attribute  $X_i$ . The x axis refers to



number  $n$ , where the  $y$  axis is the computation time. Four series are plotted for different attribute domain sizes  $m \in \{12, 22, 32, 42\}$ .

For each series, the computation time grows exponentially with the number  $n$  of parallel activities. On the other hand, the size of the input domain is roughly  $m^n$  and, hence, grows exponentially with the number  $n$  of parallel activities. Indeed, each of  $n$  attributes  $X_1, \dots, X_n$  should be assigned one out of  $m$  values.

To sum up, the experiments prove that the computation time is upper bounded by a polynomial expression in the size of the input. To have a more precise estimation, we have plotted a different graph in which the  $x$  axis is the domain size and  $y$  is the computation time. The graph is shown in Figure 4b where the dotted line delineates the regression curve that better represents the trend. For this example, the actual computation time to find a solution grows as a square root of the input domain size. This sub-linear trend demonstrates that, in practical cases, the time to find an optimal alignment is only relatively affected by the number of values a certain process attribute can be given. This remarkable result is certainly related to the goodness of the employed heuristic function. In the worst case, the theoretical complexity remains exponential in the size of the domain. But, in practice, the heuristic allows the algorithm to significantly cut the number of search-space nodes to visit and, hence, the computation time to find a solution. As a matter of fact, Figure 4c shows the number of visited nodes in case of 3 parallel activities and for different values of  $m$ . In particular, we compare such a number in the case both the heuristic is used and is unused: the heuristics roughly instructs the algorithm to only visit a logarithmic number of nodes with respect to the case when the heuristic is not used.

*Visualization of the Results in the Operationalization as ProM plug-in.* We conclude this section by showing the actual operationalization as ProM plug-in in a scenarios in which we want to check the conformance of a given log against the process of Example 1. The log contains one perfectly-fitting trace and other trace with different problems. Figure 5 illustrates how the conformance-checking results are visualized: the optimal alignment of each log trace is shown as a sequence of triangles, each representing a move in the process and/or in the log. The triangle colors represent the alignment type. The green and white color identify moves in both with the same attribute assignment or with a different one; yellow and purple report moves only in the log or in the process, respectively. When the user passes over a triangle with the mouse, the plug-in shows the execution step(s) associated to the move. The value near to every trace is the *fitness value* of the trace, i.e. a value between 0 and 1 which quantifies the quality of the alignment. Fitness value 1 identifies the perfect alignment. Conversely, a fitness value 0 pinpoints the alignment with the largest possible cost, which typically only consists by moves in log and moves in process. Interested readers can refer to [2] where fitness values are computed in the same way. At the bottom, a table shows some statistics on the attribute assignments in the moves present in the optimal alignments shown in the upper part of the screen. The second column highlights the percentage of log steps that do not provide assignment. The last two columns report the average and the standard deviation of the difference of the values assigned to attributes in the moves. We use the hamming distance to compute the string differences and, in case of boolean attributes, we consider true as value 1 and false as value 0.





**Fig. 5.** A screenshot of the ProM plug-in: how optimal alignments are visualized and what statistics are available on the process attributes

## 6 Conclusion

Process mining can be seen as the “missing link” between data mining and business process management. Although process discovery attracted the lion’s share of attention, conformance checking is at least as important. It is vital to relate process models (hand-made or discovered) to event logs. First of all, it may be used to audit processes to see whether reality conforms to some normative or descriptive model [15]. Deviations may point to fraud, inefficiencies, and poorly designed or outdated procedures. Second, conformance checking can be used to evaluate the performance of a process discovery technique. Finally, the alignment between model and log may be used for performance analysis, e.g., detecting bottlenecks [1].

Existing conformance checking techniques focus on the control flow thereby ignoring the other perspectives (data and resources). This paper presents a technique that takes data and resources into account when checking for process conformance. The proposed heuristics-based approach seems extremely promising since it allows for cutting out a significant part of the search space during the analysis. As a matter of fact, the computation time seems to be sub-linear, at least for the example used during the experiments.

Of course, a larger set of experiments with different processes is needed to verify our findings. Moreover, the absolute value of the computation time is still relatively high and that seems to be mostly related to the parsing of the guard expressions to determine the node successors in the search space. The parsing operations approximately take 70% of the overall computation time: we are currently investigating how to reduce the number of guards to be evaluated, along with integrating a more efficient parser.

**Acknowledgements.** The research leading to these results has received funding from the European Community’s Seventh Framework Programme FP7/2007-2013 under grant agreement n° 257593 (ACSI).

## References

1. van der Aalst, W.M.P.: *Process Mining - Discovery, Conformance and Enhancement of Business Processes*. Springer (2011)
2. van der Aalst, W., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(2), 182–192 (2012)
3. Rozinat, A., van der Aalst, W.: Conformance Checking of Processes Based on Monitoring Real Behavior. *Information Systems* 33, 64–95 (2008)
4. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: *Big data: The next frontier for innovation, competition, and productivity*. Technical report, McKinsey Global Institute (MGI) (May 2011)
5. van der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., de Leoni, M., Delias, P., van Dongen, B.F., Dumas, M., Dustdar, S., Fahland, D., Ferreira, D.R., Gaaloul, W., van Geffen, F., Goel, S., Günther, C., Guzzo, A., Harmon, P., ter Hofstede, A., Hoogland, J., Ingvaldsen, J.E., Kato, K., Kuhn, R., Kumar, A., La Rosa, M., Maggi, F., Malerba, D., Mans, R.S., Manuel, A., McCreesh, M., Mello, P., Mendling, J., Montali, M., Motahari-Nezhad, H.R., zur Muehlen, M., Munoz-Gama, J., Pontieri, L., Ribeiro, J., Rozinat, A., Seguel Pérez, H., Seguel Pérez, R., Sepúlveda, M., Sinur, J., Soffer, P., Song, M., Sperduti, A., Stilo, G., Stoel, C., Swenson, K., Talamo, M., Tan, W., Turner, C., Vanthienen, J., Varvaressos, G., Verbeek, E., Verdonk, M., Vigo, R., Wang, J., Weber, B., Weidlich, M., Weijters, T., Wen, L., Westergaard, M., Wynn, M.: *Process Mining Manifesto*. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) *BPM Workshops 2011, Part I. LNBP*, vol. 99, pp. 169–194. Springer, Heidelberg (2012)
6. Weijters, A., van der Aalst, W., de Medeiros, A.A.: *Process Mining with the Heuristics Miner-algorithm*. Technical report, Eindhoven University of Technology, Eindhoven, BETA Working Paper Series, WP 166 (2006)
7. de Medeiros, A.A., Weijters, A., van der Aalst, W.: Genetic Process Mining: an Experimental Evaluation. *Data Mining and Knowledge Discovery* 14, 245–304 (2007)
8. Adriansyah, A., van Dongen, B.F., van der Aalst, W.M.P.: Towards Robust Conformance Checking. In: Muehlen, M.z., Su, J. (eds.) *BPM 2010 Workshops. LNBP*, vol. 66, pp. 122–133. Springer, Heidelberg (2011)
9. Weidlich, M., Polyvyanyy, A., Desai, N., Mendling, J.: Process Compliance Measurement Based on Behavioural Profiles. In: Pernici, B. (ed.) *CAiSE 2010. LNCS*, vol. 6051, pp. 499–514. Springer, Heidelberg (2010)
10. Rozinat, A., Veloso, M., van der Aalst, W.: Using hidden markov models to evaluate the quality of discovered process models. Technical report, BPM Center Report BPM-08-10 (2008)
11. Cook, J., Wolf, A.: Software Process Validation: Quantitatively Measuring the Correspondence of a Process to a Model. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 8, 147–176 (1999)
12. Adriansyah, A., van Dongen, B., van der Aalst, W.: Conformance Checking Using Cost-Based Fitness Analysis. In: *IEEE International Enterprise Distributed Object Computing Conference*, pp. 55–64. IEEE Computer Society (2011)
13. Dechter, R., Pearl, J.: Generalized best-first search strategies and the optimality of A\*. *Journal of the ACM (JACM)* 32, 505–536 (1985)
14. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: XES, XESame, and ProM 6. In: Soffer, P., Proper, E. (eds.) *CAiSE Forum 2010. LNBP*, vol. 72, pp. 60–75. Springer, Heidelberg (2011)
15. van der Aalst, W.M.P., van Hee, K., van der Werf, J.M.E.M., Verdonk, M.: Auditing 2.0: Using Process Mining to Support Tomorrow’s Auditor. *IEEE Computer* 43(3), 90–93 (2010)

# An Approach for Consistent Delegation in Process-Aware Information Systems

Sigrid Schefer-Wenzl, Mark Strembeck, and Anne Baumgrass

Institute for Information Systems and New Media  
Vienna University of Economics and Business (WU Vienna), Austria  
firstname.lastname@wu.ac.at

**Abstract.** Delegation is an important concept to increase flexibility in authorization and obligation management. Due to the complexity of potential delegation relations, there is a strong need to systematically check the consistency of all delegation assignments. In this paper, we discuss the detection of delegation conflicts based on the formal definitions of a model that supports the delegation of roles, tasks, and duties in a business process context.

**Keywords:** Access Control, Business Processes, Delegation, RBAC.

## 1 Introduction

A business process includes a set of tasks which are performed to reach certain corporate goals. To support the secure execution of a business process, subjects participating in a particular process instance must own the permissions that are needed to execute the corresponding tasks (see, e.g., [17]). In recent years, Role-Based Access Control (RBAC) [7,9] has developed into a de facto standard for access control. In RBAC, roles are used to model different job positions and responsibilities within an organization and/or information system. Permissions are assigned to roles according to the tasks each role has to accomplish. The roles are then assigned to human users according to their respective work profile [15]. Roles are also used as an abstract concept for delegation [5,18] or for the assignment of duties defined via obligations [11,21].

Authorization policies define a subject's permissions, while obligation policies define a subject's duties (see, e.g., [3]). Delegation provides a mechanism to increase flexibility in authorization and obligation management. In essence, a subject can delegate tasks, duties, or roles to another subject [11]. Subsequently, the subject receiving the delegation (the delegatee) will act on behalf of the delegating subject (the delegator). While delegation authorizes subjects to perform tasks they usually are not allowed to perform, authorization constraints, such as mutual-exclusion (ME) and binding constraints, restrict which subject is allowed to execute a particular task (see, e.g., [16,17,19]). In process-aware information systems, ME constraints enforce conflict of interest policies. Conflict of interest arises as a result of the simultaneous assignment of two mutually exclusive tasks or roles to the same subject. In contrast to ME constraints, binding constraints

define that bound tasks must be executed by the same subject or role. The immanent complexity of delegations is a central problem in process-aware information systems (see, e.g., [4,10]). Thus, when delegating tasks, roles, or duties, design-time and run-time checks need to ensure the consistency of the corresponding RBAC model including mutual-exclusion and binding constraints. In [13,16], we provide a set of algorithms that check and ensure the consistency of process-related RBAC models without addressing delegation aspects.

The main contribution of this paper is the consideration of delegations when checking and ensuring the consistency of process-related RBAC models. In particular, we integrate the formal definitions of our delegation model into process-related RBAC models [17]. These definitions are based on several existing, well-known delegation models and are the basis for the algorithms presented in this paper. The algorithms systematically detect potential conflicts when delegating roles, tasks, and duties at design- and run-time. For this purpose, we take the conflicts identified in [13,16] as a starting point.

The remainder of this paper is structured as follows. In Section 2, we introduce relevant terms and present the formal definitions of a process-related RBAC delegation model. Section 3 provides algorithms to detect potential delegation conflicts to ensure the consistency of a process-related RBAC delegation model. Section 4 discusses related work and Section 5 concludes the paper.

## 2 Process-Related RBAC Delegation Models

In our process-related RBAC delegation model, roles, tasks, and associated duties are delegatable. Each *task* in an IT-supported workflow (such as negotiating a contract) is typically associated with certain access operations (e.g., to sign the contract). Thus, a subject participating in a workflow must be authorized to perform the tasks needed to complete the process (see, e.g., [17]). In organizational contexts, tasks can be associated with duties. Each *duty* defines an action that must be performed by a certain subject in order to comply with legal or organizational regulations (see, e.g., [3,12]). A *subject* may either be a human user or a software-based system. In RBAC, a *role* is a subject abstraction containing the tasks and duties of a certain subject-type.

In the context of RBAC, several delegation approaches use the concept of *delegation roles* (see, e.g., [8,14,20]). In our delegation model, a delegation role is created by the *delegator* and comprises a *set of delegated tasks and duties* (similar to [20]). Hereby, each duty is associated with a certain task [12]. A delegator can delegate all or a subset of his/her delegatable tasks, duties, or roles by assigning them to a delegation role. Subsequently, delegation roles are assigned to delegates and can either be defined for temporary or for permanent delegation (see, e.g., [2,20]). By default, delegation roles are permanent which means they authorize the delegatee to perform the delegated tasks and duties in all instances of a process. In contrast, a temporary delegation role authorizes the delegatee to perform the delegated tasks and duties only in particular process instances. Moreover, we support single- and multi-step delegation (see, e.g., [2,18]).

In single-step delegation, a delegated task, duty, or role cannot be delegated further by the delegatee. Multi-step delegation allows a delegatee to further delegate the delegated tasks, duties, and roles. In general, delegation roles and all assignments to delegation roles are managed by the delegating subject. All other roles are called *regular roles* and are usually managed by the organization’s security officer. Fig. 1 shows a class diagram that depicts the elements of the RBAC delegation model (see Definition 1).

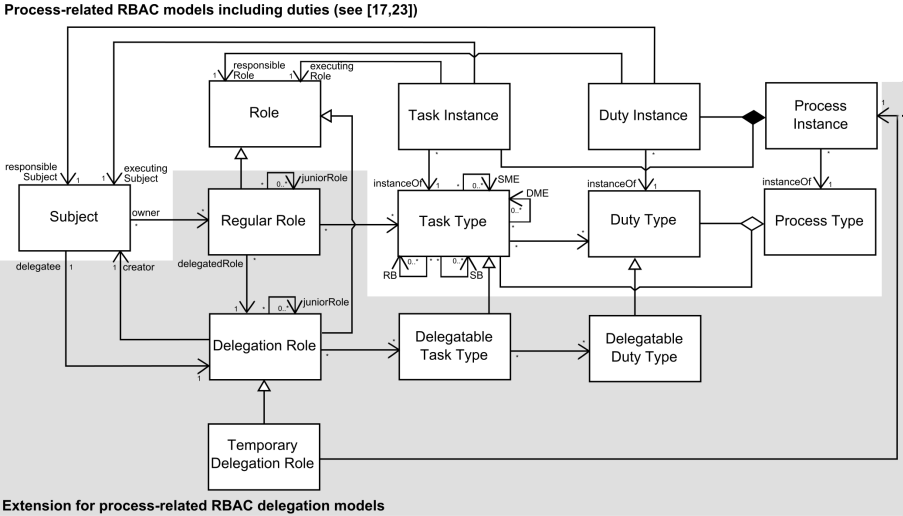


Fig. 1. Main elements of process-related RBAC delegation models

Furthermore, different kinds of authorization constraints can be defined to restrict which subjects are allowed to execute a particular task or duty (see, e.g., [17,19]). In this paper, we focus on static mutual exclusion (SME), dynamic mutual exclusion (DME), subject-binding (SB), and role-binding (RB) constraints. A SME constraint defines that two statically mutual exclusive tasks must never be *assigned* to the same subject. In turn, a DME constraint defines that two dynamically mutual exclusive tasks must never be *executed* by the same subject in the *same process instance*. A SB constraint defines that two bound tasks must be performed by the same individual within the same process instance. A RB constraint defines that bound tasks must be performed by members of the same role, but not necessarily by the same individual. To ensure proper delegation, authorization constraints must be considered when delegating tasks, duties, and roles (see Section 3). For example, a delegation assignment must not authorize the delegatee to perform two SME tasks.

Definition 1 formally specifies the essential elements and their basic interrelations in a metamodel for process-related RBAC delegation models (see Fig. 1).

**Definition 1. (Process-Related RBAC Delegation Model).** Let  $PRDM = (E, Q, D, DL)$  be a Process-Related RBAC Delegation Model, where  $E$  refers to the pairwise disjoint sets of the metamodel,  $Q$  to mappings that establish relationships,  $D$  to binding and mutual-exclusion constraints, and  $DL$  to mappings for delegation policies.

The sets  $E$  of the Process-Related RBAC Delegation Model are:

- An element of  $S$  is called Subject.  $S \neq \emptyset$ .
- An element of  $R$  is called Role.  $R \neq \emptyset$ .
- An element of  $RR$  is called Regular Role.  $RR \subseteq R$ .
- An element of  $DR$  is called Delegation Role.  $DR \subseteq R$
- An element of  $DRT$  is called Temporary Delegation Role.  $DRT \subseteq DR$ .
- An element of  $P_T$  is called Process Type.  $P_T \neq \emptyset$ .
- An element of  $P_I$  is called Process Instance.  $P_I \neq \emptyset$ .
- An element of  $T_T$  is called Task Type.  $T_T \neq \emptyset$ .
- An element of  $DT_T$  is called Delegatable Task Type.  $DT_T \subseteq T_T$ .
- An element of  $T_I$  is called Task Instance.
- An element of  $DU_T$  is called Duty Type.
- An element of  $DDU_T$  is called Delegatable Duty Type.  $DDU_T \subseteq DU_T$ .
- An element of  $DU_I$  is called Duty Instance.

For the mappings of the Process-Related RBAC Model ( $Q, D$ ) see [17]. Below, we define additional mappings for delegation:  $DL = rrh \cup drh \cup creator \cup drpi \cup trra \cup trdel \cup dta \cup rrsa \cup rsdel \cup dui \cup res \cup rer$  ( $\mathcal{P}$  refers to the power set):

1. Roles  $R$  are partitioned into regular roles and delegation roles. In RBAC, roles can be arranged in a role-hierarchy, where senior-roles inherit the permissions from their junior-roles. To avoid invalid permission inheritance, the regular role-hierarchy consists of regular roles only. If a model uses process-related RBAC delegation, this mapping replaces the role-hierarchy mapping  $rh$  in [17]: The mapping  $rrh : RR \mapsto \mathcal{P}(RR)$  is called **regular role-hierarchy**. For  $rrh(r_s) = RR_j$ , we call  $r_s \in RR$  senior regular role and  $RR_j \subseteq RR$  the set of direct junior regular roles. The transitive closure  $rrh^*$  defines the inheritance in the role-hierarchy such that  $rrh^*(r_s) = RR_{j^*}$  includes all direct and transitive regular junior-roles that the senior-role  $r_s$  inherits from. The regular role-hierarchy is cycle-free, i.e. for each  $r \in RR : rrh^*(r) \cap r = \emptyset$ .
2. Delegation roles can be arranged in a delegation role-hierarchy via role-to-role delegation. Note that each delegation role may have junior regular roles or junior delegation roles (see, e.g., [20]). However, delegation roles must not have senior regular roles to avoid invalid permission inheritance in the regular role hierarchy: The mapping  $drh : DR \mapsto \mathcal{P}(R)$  is called **delegation role-hierarchy**. For  $drh(dr_s) = R_j$ , we call  $dr_s \in DR$  senior delegation role and  $R_j \subseteq R$  the set of direct junior-roles. The transitive closure  $drh^*$  defines the inheritance in the role-hierarchy such that  $drh^*(dr_s) = R_{j^*}$  includes all direct and transitive junior-roles that the senior-role  $dr_s$  inherits from. The delegation role-hierarchy is cycle-free, i.e. for each  $r \in R : drh^*(r) \cap r = \emptyset$ .

3. Each subject can create an arbitrary number of delegation roles. Subsequently, the creator will act as the delegator of its delegation roles: *The mapping  $\text{creator}(dr) : DR \mapsto S$  is called **delegation role creator**. For  $\text{creator}(dr) = s$ , we call  $dr \in DR$  delegation role and  $s \in S$  the creator of this delegation role.*
4. Each delegation role can be specified either for permanent or for temporary delegation. By default, a delegation role is permanent and is valid for all process types. In case of temporary delegation, a temporary delegation role is only valid for particular process instances: *The mapping  $\text{drpi} : DRT \mapsto \mathcal{P}(P_I)$  is called **delegation role-to-process assignment**. For  $\text{drpi}(drt) = P_{drt}$ , we call  $drt \in DRT$  temporary delegation role, and  $P_{drt} \subseteq P_I$  the set of process instances.*
5. Task types are assigned to regular roles to define the permissions of the corresponding role. If a model uses process-related RBAC delegation, this mapping replaces the task-to-role assignment mapping  $\text{tra}$  in [17]: *The mapping  $\text{trra} : RR \mapsto \mathcal{P}(T_T)$  is called **task-to-regular role assignment**. For  $\text{trra}(r) = T_r$ , we call  $r \in RR$  regular role and  $T_r \subseteq T_T$  is called the set of tasks assigned to  $r$ . The mapping  $\text{trra}^{-1} : T_T \mapsto \mathcal{P}(RR)$  returns the set of regular roles a particular task is assigned to.*
6. Task types can be defined as being delegatable. Only delegatable tasks can be assigned to delegation roles. Thus, a subject can delegate a task by assigning this task to a delegation role: *The mapping  $\text{trdel} : DR \mapsto \mathcal{P}(DT_T)$  is called **task-to-role delegation**. For  $\text{trdel}(dr) = DT_{dr}$ , we call  $dr \in DR$  delegation role and  $DT_{dr} \subseteq DT_T$  is called the set of delegated tasks assigned to  $dr$ . The mapping  $\text{trdel}^{-1} : DT_T \mapsto \mathcal{P}(DR)$  returns the set of delegation roles a particular delegatable task is assigned to.*
7. Further,  $\text{trra}$  and  $\text{trdel}$  imply a mapping **task ownership**  $\text{town} : R \mapsto \mathcal{P}(T_T)$  to determine all tasks that are assigned to a particular role. If a model uses process-related RBAC delegation, this mapping replaces the  $\text{town}$ -mapping from [17]: *For each  $r \in R$ , the tasks inherited from its junior roles are included, i.e.  $\text{town}(r) = \text{town}_{rrh}(r) \cup \text{town}_{drh}(r)$ , where  $\text{town}_{rrh}(r) = \bigcup_{r_{inh} \in rrrh^*(r)} \text{trra}(r_{inh}) \cup \text{trra}(r)$  and  $\text{town}_{drh}(r) = \bigcup_{r_{inh} \in drh^*(r)} \text{trdel}(r_{inh}) \cup \text{trdel}(r)$ .*
8. A duty defines an action that must be performed by a certain subject. In a business process context, each duty is associated with a task [12]: *The mapping  $\text{dta} : T_T \mapsto \mathcal{P}(DU_T)$  is called **duty-to-task assignment**. For  $\text{dta}(t) = DU_x$ , we call  $t \in T_T$  task type and  $DU_x \subseteq DU_T$  is called the set of duties assigned to this task type.*
9. Delegatable tasks can only be delegated, if all associated duties are also delegatable:  $\forall t_x \in \text{trdel}(dr) : \forall du \in \text{dta}(t_x) : du \in DDU_T$
10. Regular roles are assigned to subjects. Thereby, subjects acquire the rights to execute the corresponding tasks and duties. If a model uses process-related RBAC delegation, this mapping replaces the role-to-subject assignment mapping  $\text{rsa}$  in [17]: *The mapping  $\text{rrsa} : S \mapsto \mathcal{P}(RR)$  is called **regular role-to-subject assignment**. For  $\text{rrsa}(s) = RR_s$ , we call  $s \in S$*

subject and  $RR_s \in RR$  the set of regular roles owned by  $s$ . The mapping  $rrsa^{-1} : RR \mapsto \mathcal{P}(S)$  returns all subjects assigned to a regular role.

11. Delegation roles are assigned to delegates who are subsequently authorized and responsible to perform the corresponding delegated tasks and duties: The mapping  $rsdel : S \mapsto \mathcal{P}(DR)$  is called **role-to-subject delegation**. For  $rsdel(s) = DR_s$ , we call  $s \in S$  delegatee and  $DR_s \in DR$  the set of delegation roles owned by  $s$ . The mapping  $rsdel^{-1} : DR \mapsto \mathcal{P}(S)$  returns all delegates assigned to a delegation role.
12. Further,  $rrsa$  and  $rsdel$  imply a mapping **role ownership**  $rown : S \mapsto \mathcal{P}(R)$  to determine all roles that are assigned to a particular subject. If a model uses process-related RBAC delegation, this mapping replaces the  $rown$ -mapping from [17]: For each  $s \in S$ , all inherited roles are included, i.e.  $rown(s) = rown_{rrh}(s) \cup rown_{drh}(s)$ , where  $rown_{rrh}(s) = \bigcup_{r \in rrsa(s)} rrh^*(r) \cup rrsa(s)$  and  $rown_{drh}(s) = \bigcup_{r \in rsdel(s)} drh^*(r) \cup rsdel(s)$ .
13. For each task type, we can create an arbitrary number of respective task instances via the **task instantiation** mapping  $ti$  [17]. Similarly, each duty type is instantiated by a number of duty instances: The mapping  $dui : (DU_T \times P_I) \mapsto \mathcal{P}(DU_I)$  is called **duty instantiation**. For  $dui(du_T, p_I) = DU_i$ , we call  $DU_i \subseteq DU_I$  set of duty instances,  $du_T \in DU_T$  is called duty type and  $p_I \in P_I$  is called process instance.
14. The **executing-subject** mapping  $es$  returns the subject executing a particular task instance [17]. The subject responsible for discharging a duty is called the responsible subject of this duty instance: The mapping  $res : DU_I \mapsto S$  is called **responsible-subject** mapping. For  $res(du) = s$ , we call  $s \in S$  the responsible subject and  $du \in DU_I$  is called duty instance.
15. Within the same process instance, a subject executing a task is also responsible for discharging all associated duties:  $\forall du \in dta(t_1), p_i \in P_I : \forall t_x \in ti(t_1, p_i), du_x \in dui(du, p_i) : es(t_x) = res(du_x)$
16. The **executing-role** mapping  $er$  returns the role executing a particular task instance [17]. The **active-role** mapping  $ar$  returns the role a subject has currently activated [16]. The role being responsible for actually discharging a certain duty instance is called the responsible-role: The mapping  $rer : DU_I \mapsto R$  is called **responsible-role** mapping. For  $rer(du) = r$ , we call  $r \in R$  the responsible role and  $du \in DU_I$  is called discharged duty instance.
17. Further, we allow the definition of subject-binding, role-binding, static mutual exclusion, and dynamic mutual exclusion constraints on task types. Related consistency requirements are specified in [17]: The mapping  $sb : T_T \mapsto \mathcal{P}(T_T)$  is called **subject-binding**. For  $sb(t_1) = T_{sb}$ , we call  $t_1$  the subject binding task and  $T_{sb} \subseteq T_T$  the set of subject-bound tasks. The mapping  $rb : T_T \mapsto \mathcal{P}(T_T)$  is called **role-binding**. For  $rb(t_1) = T_{rb}$ , we call  $t_1$  the role binding task and  $T_{rb} \subseteq T_T$  the set of role-bound tasks. The mapping  $sme : T_T \mapsto \mathcal{P}(T_T)$  is called **static mutual exclusion**. For  $sme(t_1) = T_{sme}$  with  $T_{sme} \subseteq T_T$ , we call each pair  $t_1$  and  $t_x \in T_{sme}$  statically mutual exclusive tasks. The mapping  $dme : T_T \mapsto \mathcal{P}(T_T)$  is called **dynamic mutual exclusion**. For  $dme(t_1) = T_{dme}$  with  $T_{dme} \subseteq T_T$ , we call each pair  $t_1$  and  $t_x \in T_{dme}$  dynamically mutual exclusive tasks.



### 3 Detecting Delegation Conflicts

When delegating tasks, duties, or roles several conflicts may occur. In [13,16], we detect conflicts of process-related RBAC models at design-time and run-time. In this paper, we provide additional algorithms to detect delegation conflicts. Algorithms [1-3] check the design-time consistency of a process-related RBAC delegation model when defining a task-to-role, role-to-role, or role-to-subject delegation relation. Algorithm [4] checks the consistency of a process-related RBAC delegation model at run-time. All other conflicts that can potentially occur at design- or run-time are addressed by the algorithms presented in [13,16].

**Algorithm 1.** Check if it is allowed to delegate a task type to a delegation role.

*Input:*  $task_x \in T_T, drole_y \in DR, delegator \in S$

- 1: if  $delegator \neq creator(drole_y)$  then return false
- 2: if  $task_x \notin DT_T$  then return false
- 3: if  $\exists duty_x \in dta(task_x) \mid duty_x \notin DDU_T$  then return false
- 4: if  $\nexists r \in rown(delegator) \mid task_x \in town(r) \wedge r \in RR$  then return false
- 5: if  $\exists task_y \in town(drole_y) \mid task_y \in sme(task_x)$  then return false
- 6: if  $\exists role_z \in allSeniorRoles(drole_y) \mid task_z \in town(role_z) \wedge$   
7:  $task_z \in sme(task_x)$  then return false
- 8: if  $\exists s \in S \mid drole_y \in rown(s) \wedge role_z \in rown(s) \wedge$   
9:  $task_z \in town(role_z) \wedge task_z \in sme(task_x)$  then return false
- 10: if  $\exists task_y \in sb(task_x) \mid task_y \notin DT_T$  then return false
- 11: if  $\exists task_y \in rb(task_x) \mid task_y \notin DT_T$  then return false
- 12: if  $\exists task_y \in sb(task_x) \mid duty_y \in dta(task_y) \wedge duty_y \notin DDU_T$  then return false
- 13: if  $\exists task_y \in rb(task_x) \mid duty_y \in dta(task_y) \wedge duty_y \notin DDU_T$  then return false
- 14: return true

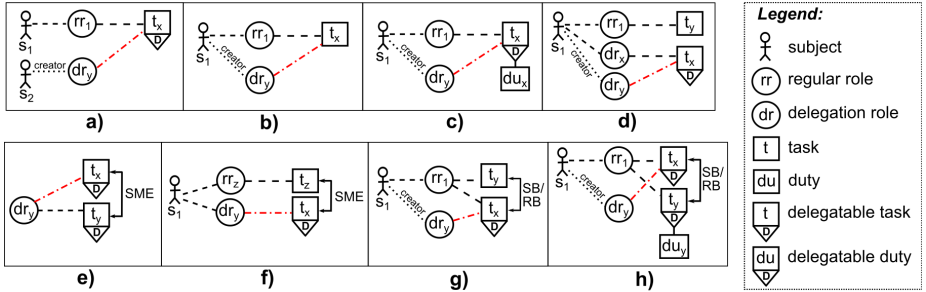


Fig. 2. Delegation conflicts

Only the creator of a delegation role can delegate to it and assign delegates. Thus, Algorithm [1], line 1 returns false if a subject tries to delegate to a delegation role which he/she has not created. For example, in Fig. [2]a, subject  $s_1$  tries to delegate task  $t_x$  to delegation role  $dr_y$ . Task  $t_x$  is delegatable which is visualized in Fig. [2] by a triangle attached to the task-symbol including the letter D. However,  $s_1$  is not the creator of  $dr_y$  and thus  $s_1$  cannot delegate to it.

Next, Algorithm [1](#), line 2 checks if a subject tries to delegate a task which is not delegatable. In Fig. [2b](#), task  $t_x$  cannot be delegated to delegation role  $dr_y$ , because  $t_x$  is not delegatable. Afterwards, line 3 checks if a subject tries to delegate a task which is associated with a non-delegatable duty. Duties always need to be discharged by the subject executing the corresponding task. Thus, if a task is delegated, the corresponding duty also needs to be delegatable. In Fig. [2c](#), task  $t_x$  can not be delegated to delegation role  $dr_y$ , because the duty  $du_x$  associated to  $t_x$  is not delegatable.

Algorithm [1](#), line 4 returns false if a subject tries to delegate a task which he/she is not assigned to via its (regular) role ownership assignments. If single-step delegation is preferred, the subject can only delegate tasks and duties which he/she owns directly or transitively via a regular role. This is because single-step delegation does not allow to further delegate a delegated task. In Fig. [2d](#), subject  $s_1$  tries to delegate task  $t_x$  to its delegation role  $dr_y$ . However, none of the *regular* roles owned by  $s_1$  is assigned to  $t_x$ . Thus,  $s_1$  cannot delegate  $t_x$  to  $dr_y$ . In case of multi-step delegation, a subject can delegate all of his/her tasks and duties. For this purpose, we need to change the condition  $r \in RR$  in Algorithm [1](#), line 4 to  $r \in R$ . Subsequently, a subject can delegate tasks and duties which he/she owns directly or transitively via its regular or delegation role memberships.

Moreover, it is not possible to delegate a task if this delegation would result in the assignment of two SME tasks to the same role or subject (see Algorithm [1](#), lines 5-9). Fig. [2e](#) depicts an example where a delegation role  $dr_y$  owns a task  $t_y$  which is defined as SME to another task  $t_x$ . Thus,  $t_x$  cannot be delegated to  $dr_y$ . Otherwise,  $dr_y$  would subsequently own two SME tasks. Fig. [2f](#) shows an example, where the delegation of task  $t_x$  to the delegation role  $dr_y$  is not allowed because  $s_1$  would then be authorized to perform the two SME tasks  $t_z$  and  $t_x$ .

If a subject tries to delegate a task which has a subject-binding to one or more non-delegatable task(s), Algorithm [1](#), line 10 returns false. This is because subject-bound tasks always have to be performed by the same subject. Thus, if a task is delegated, all subject-bound tasks also need to be assigned to the same delegation role. Otherwise, the SB constraint cannot be fulfilled. In Fig. [2g](#), a SB constraint is defined on  $t_x$  and  $t_y$ . Therefore, the subject performing  $t_x$  also has to perform  $t_y$ . When delegating  $t_x$  to  $dr_y$  Algorithm [1](#) returns false, because  $t_y$  is not delegatable. However, to fulfill the SB constraint, both tasks need to be delegated to  $dr_y$ . Similarly, Algorithm [1](#), line 11 returns false if a subject tries to delegate a task which has a role-binding to one or more non-delegatable task(s). Thus, if a task is delegated, all role-bound tasks also need to be assigned to the same delegation role.

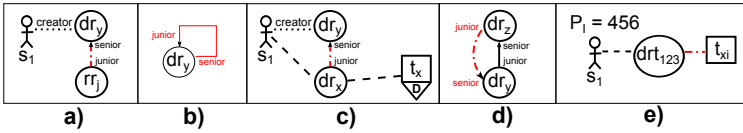
Furthermore, a subject cannot delegate a task which has a subject-binding to other tasks, if one of the subject-bound tasks is associated with a non-delegatable duty. In Fig. [2h](#), a SB constraint is defined on  $t_x$  and  $t_y$ . Moreover,  $t_y$  is associated with a duty  $du_y$ . If subject  $s_1$  tries to delegate  $t_x$  to  $dr_y$ , it also has to delegate all subject-bound tasks and associated duties. In this example,  $du_y$  is not delegatable. Thus, Algorithm [1](#), line 12 returns false. Similarly, if a subject tries to delegate a task which has a role-binding to other tasks, Algorithm [1](#), line 13 returns false, if one of the role-bound tasks is associated with a non-delegatable duty. If none of the above

checks returns false, Algorithm 2 finally reaches line 14 and returns true – meaning that it is allowed to delegate a particular task type to a certain delegation role.

**Algorithm 2.** *Check if it is allowed to delegate a role to a delegation role.*

**Input:**  $junior \in R, senior \in DR, delegator \in S$

- 1: **if**  $delegator \neq creator(senior)$  **then return false**
- 2: **if**  $junior \notin rown(delegator)$  **then return false**
- 3: **if**  $junior == senior$  **then return false**
- 4: **if**  $\exists task_x \in town(junior) \mid task_x \notin DT$  **then return false**
- 5: **if**  $\exists task_x \in town(junior) \mid duty_x \in dta(task_x) \wedge$
- 6:  $duty_x \notin DDU_T$  **then return false**
- 7: **if**  $junior \in DR$  **then**  $\exists r \in rown(delegator) \mid task_x \in town(junior) \wedge$
- 8:  $task_x \in town(r) \wedge r \in RR$  **else return false**
- 9: **if**  $senior \in drh^*(junior)$  **then return false**
- 10: **if**  $\exists task_j \in town(junior) \mid task_s \in town(senior) \wedge$
- 11:  $task_j \in sme(task_s)$  **then return false**
- 12: **if**  $\exists role_x \in allSeniorRoles(senior) \mid task_x \in town(role_x) \wedge$
- 13:  $task_j \in town(junior) \wedge task_x \in sme(task_j)$  **then return false**
- 14: **if**  $\exists s \in S \mid senior \in rown(s) \wedge role_x \in rown(s) \wedge task_x \in town(role_x) \wedge$
- 15:  $task_j \in town(junior) \wedge task_x \in sme(task_j)$  **then return false**
- 16: **if**  $\exists task_x \in town(junior) \mid task_y \in sb(task_x) \wedge task_y \notin DT_T$  **then return false**
- 17: **if**  $\exists task_x \in town(junior) \mid task_y \in sb(task_x) \wedge duty_y \in dta(task_y) \wedge$
- 18:  $duty_y \notin DDU_T$  **then return false**
- 19: **return true**



**Fig. 3.** Delegation conflicts

Algorithm 2 first checks if the delegator of a role is the creator of the corresponding delegation role. Subsequently, line 2 checks if a subject tries to delegate a role which he/she is not assigned to. In Fig. 3a, subject  $s_1$  tries to delegate the regular role  $rr_j$  to its delegation role  $dr_y$  by assigning  $rr_j$  as junior-role of  $dr_y$ . However,  $s_1$  is not assigned to  $rr_j$  and thus  $s_1$  cannot delegate  $rr_j$ .

Next, Algorithm 2, line 3 returns false when delegating a role to itself. In general, a role cannot be its own junior-role (see Fig. 3b and [16, 17]). Algorithm 2, line 4 checks if the role which is to be delegated only contains delegatable tasks. Similarly, lines 5-6 check if all duties associated to the tasks of the corresponding role are delegatable. If either tasks or duties assigned to the role are not delegatable, Algorithm 2 returns false. Algorithm 2, lines 7-8 check if a subject tries to delegate a delegation role owning a task which the delegator is not assigned to via its *regular* role memberships (single-step delegation). Thus, a subject can only delegate tasks and duties which he/she owns directly or transitively via a regular role (see Figure 3c). In case of multi-step delegation, we can omit this check. Moreover, a role-hierarchy must not include a cycle because all roles

within such a cyclic inheritance relation would own the same permissions which would render the respective part of the role-hierarchy redundant. Line 9 returns false if a subject tries to delegate a role to a delegation role which is already defined as its senior-role (see Fig. 3f and 16,17).

Afterwards, Algorithm 2, lines 10-15 prevent that a role-to-role delegation would result in the assignment of two SME tasks to the same role or subject. In particular, this conflict occurs if the potential senior-role owns a task which is SME to one of the tasks owned by the potential junior-role. Subsequently, if a subject tries to delegate a role owning a task which has a subject-binding to one or more non-delegatable task(s), Algorithm 2, line 16 returns false. This is because subject-bound tasks always have to be performed by the same subject. In case a subject tries to delegate a role owning a task which has a subject-binding to other tasks, Algorithm 2, line 18 returns false, if one of the subject-bound tasks is associated with a non-delegatable duty. If none of the above checks returns false, Algorithm 2 finally reaches line 19 and returns true – meaning that it is allowed to delegate a particular role to a certain delegation role.

**Algorithm 3.** *Check if it is allowed to assign a particular delegation role to a certain delegatee.*

**Input:**  $drole_x \in DR, delegatee, delegator \in S$   
 1: **if**  $delegator \neq creator(drole_x)$  **then return false**  
 2: **if**  $\exists role_y \in rown(delegatee) \mid task_y \in town(role_y) \wedge$   
 3:  $task_x \in town(drole_x) \wedge task_y \in sme(task_x)$  **then return false**  
 4: **return true**

Algorithm 3, line 1 returns false if the subject who wants to assign a delegation role to a particular delegatee is not the creator of this delegation role. Subsequently, we need to check if the delegatee-assignment would result in the assignment of SME tasks to the delegatee (due to other role-memberships of the delegatee). If none of the above checks returns false, Algorithm 3 finally reaches line 4 and returns true – meaning that it is allowed to assign a particular delegatee to a certain delegation role.

**Algorithm 4.** *Check if a particular task instance that is executed in a certain process instance can be allocated to a specific delegatee.*

**Input:**  $drole \in DRT, delegatee \in S, task_{type} \in T_T, process_{type} \in P_T,$   
 $process_{instance} \in pi(process_{type}), task_{instance} \in ti(task_{type}, process_{instance})$   
 1: **if**  $\exists instance_y \in ti(type_y, process_{instance}) \mid ar(delegatee) = drole \wedge$   
 2:  $process_{instance} \notin drpi(drole)$  **then return false**  
 3: **return true**

Algorithm 4, line 2 returns false if the selected subject is not allowed to execute a certain task instance because the *temporary delegation role* is not valid for the corresponding process instance. Each temporary delegation role is only valid for particular process instances (see Definition 14). In Fig. 3e, subject  $s_1$  is assigned to the temporary delegation role  $drt$ , and  $drt$  is only valid for the process instance 123. However, the actual process instance is 456. Thus,  $s_1$  is not allowed to execute the delegated tasks in this process instance. Note that this check is not required for permanent delegation roles.

## 4 Related Work

In recent years, there has been much work on various aspects of role- and permission-based delegation (see, e.g., [2,20]). Delegation in a business process/workflow context has also received considerable attention. In [1], the notion of delegation is extended to allow for conditional delegation in workflows. Different types of constraints, such as authorization constraints, are addressed in the context of delegation. The effects of some delegation operations on three workflow execution models are described in [6]. Few contributions exist which consider authorization constraints and related consistency conflicts in the context of delegation. In [14], an extension to PBDM is presented to integrate authorization constraints in permission-based delegation. [14] focuses on static separation of duty constraints and related conflicts and analyzes role-based constraints. In [4], Crampton addresses the satisfiability problem of workflows in the context of constrained delegation and provides an algorithm that determines whether to permit a delegation request. However, the algorithm does not distinguish between different conflict types. In [10], Schaad discusses delegation conflicts. In contrast to our approach, the conflicts are detected after conducting the delegation, while our algorithms detect conflicts before the delegation is performed. Thus, we aim to detect conflicts before causing an inconsistent RBAC configuration.

## 5 Conclusion

In this paper, we provide a formal metamodel for process-related RBAC delegation models. In addition, we presented generic algorithms to detect conflicts in the context of delegating tasks, duties, and roles. We also discuss the specific problem of mutual-exclusion and binding constraints in an RBAC delegation context. Note that in our approach, conflicts are detected before causing an inconsistent RBAC configuration. Thus, the application of the algorithms presented in this paper helps security engineers to prevent design- and run-time conflicts in access control models and thereby aims to ensure the continuous consistency of corresponding process-related RBAC delegation models.

## References

1. Atluri, V., Warner, J.: Supporting Conditional Delegation in Secure Workflow Management Systems. In: Proceedings of the 10th ACM Symposium on Access Control Models and Technologies, SACMAT (June 2005)
2. Barka, E., Sandhu, R.: Framework for Role-Based Delegation Models. In: Proceedings of the 16th Annual Computer Security Applications Conference, ACSAC (December 2000)
3. Cole, J., Derrick, J., Milosevic, Z., Raymond, K.: Author Obligated to Submit Paper before 4 July: Policies in an Enterprise Specification. In: Sloman, M., Lobo, J., Lupu, E.C. (eds.) POLICY 2001. LNCS, vol. 1995, pp. 1–17. Springer, Heidelberg (2001)

4. Crampton, J., Khambhammettu, H.: Delegation and Satisfiability in Workflow Systems. In: Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, SACMAT (June 2008)
5. Crampton, J., Khambhammettu, H.: Delegation in Role-Based Access Control. *International Journal of Information Security* 7(2) (2008)
6. Crampton, J., Khambhammettu, H.: On Delegation and Workflow Execution Models. In: Proceedings of the 2008 ACM Symposium on Applied Computing, SAC (March 2008)
7. Ferraiolo, D.F., Kuhn, D.R., Chandramouli, R.: *Role-Based Access Control*, 2nd edn. Artech House (2007)
8. Joshi, J.B.D., Bertino, E.: Fine-grained Role-based Delegation in Presence of the Hybrid Role Hierarchy. In: Proceedings of the 11th ACM Symposium on Access Control Models and Technologies, SACMAT (June 2006)
9. Sandhu, R., Coyne, E., Feinstein, H., Youman, C.: *Role-Based Access Control Models*. *IEEE Computer* 29(2) (1996)
10. Schaad, A.: Detecting Conflicts in a Role-Based Delegation Model. In: Proceedings of the 17th Annual Computer Security Applications Conference, ACSAC (2001)
11. Schaad, A., Moffett, J.D.: Delegation of Obligations. In: Proceedings of the 3rd International Workshop on Policies for Distributed Systems and Networks, POLICY (June 2002)
12. Schefer, S., Strembeck, M.: Modeling Process-Related Duties with Extended UML Activity and Interaction Diagrams. *Electronic Communications of the EASST* 37 (March 2011)
13. Schefer, S., Strembeck, M., Mendling, J., Baumgrass, A.: Detecting and Resolving Conflicts of Mutual-Exclusion and Binding Constraints in a Business Process Context. In: Meersman, R., Dillon, T., Herrero, P., Kumar, A., Reichert, M., Qing, L., Ooi, B.-C., Damiani, E., Schmidt, D.C., White, J., Hauswirth, M., Hitzler, P., Mohania, M. (eds.) *OTM 2011, Part I. LNCS*, vol. 7044, pp. 329–346. Springer, Heidelberg (2011)
14. Shang, Q., Wang, X.: Constraints for Permission-Based Delegations. In: Proceedings of the 8th IEEE International Conference on Computer and Information Technology Workshops, CITWORKSHOPS (2008)
15. Strembeck, M.: Scenario-Driven Role Engineering. *IEEE Security & Privacy* 8(1) (2010)
16. Strembeck, M., Mendling, J.: Generic Algorithms for Consistency Checking of Mutual-Exclusion and Binding Constraints in a Business Process Context. In: Meersman, R., Dillon, T.S., Herrero, P. (eds.) *OTM 2010, Part I. LNCS*, vol. 6426, pp. 204–221. Springer, Heidelberg (2010)
17. Strembeck, M., Mendling, J.: Modeling Process-related RBAC Models with Extended UML Activity Models. *Information and Software Technology* 53(5) (2011)
18. Wainer, J., Kumar, A., Barthelmeß, P.: DW-RBAC: A formal security model of delegation and revocation in workflow systems. *Information Systems* 32(3) (2007)
19. Warner, J., Atluri, V.: Inter-Instance Authorization Constraints for Secure Workflow Management. In: Proceedings of the 11th ACM Symposium on Access Control Models and Technologies, SACMAT (June 2006)
20. Zhang, X., Oh, S., Sandhu, R.: PBDM: A Flexible Delegation Model in RBAC. Proceedings of the 8th ACM Symposium on Access Control Models and Technologies, SACMAT (June 2003)
21. Zhao, G., Chadwick, D., Otenko, S.: Obligations for Role Based Access Control. In: Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops, AINAW (May 2007)

# Goal-Oriented Model-Driven Business Process Monitoring Using ProGoalML

Falko Koetter<sup>1</sup> and Monika Kochanowski<sup>2</sup>

<sup>1</sup> University of Stuttgart IAT, Germany  
falko.koetter@iao.fraunhofer.de

<sup>2</sup> Fraunhofer IAO, Germany  
monika.kochanowski@iao.fraunhofer.de

**Abstract.** In today's fast changing business world, the fulfillment of process goals needs to be constantly evaluated and adjusted. But processes are often carried out by systems which are not process aware. aPro is a modular architecture for business process optimization. In aPro process models can't be guaranteed to be executable but need to be monitored. In this paper, we propose a modeling language for process metrics, key performance indicators and goals and use the interchange format ProGoalML to automate creation and setup of monitoring infrastructure.

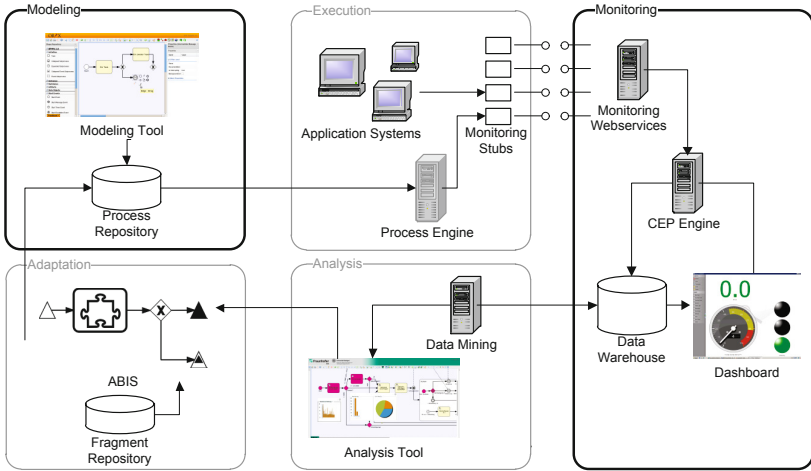
**Keywords:** business process management, process monitoring, business process goals, process adaptation, business intelligence.

## 1 Introduction

In today's commerce business processes are volatile and interconnected, causing the necessity to react to changes in a timely and correct fashion [8]. But only changes impacting the goals of the process necessitate an adjustment of the process. Thus, to assess the need for change, the goals of a process as well as their degree of fulfillment have to be known. Monitoring solutions today [21,2,3] focus on executable processes.

However, our work in the industry as well as other sources [16] [14] found executable process models to be the exception rather than the norm. Thus, business process models are often disconnected from process execution and serve only documentation purposes. While switching to executable processes is often not possible due to existing systems or lack of IT support in single process steps, there still is a need for business process monitoring and optimization [14]. Companies need a way to implement missing capabilities without abandoning existing solutions. Thus, the contribution of this work is to provide a model-driven approach for monitoring business processes which takes into account the current state of BPM adoption.

We propose aPro, a modular Architecture for business *PR*ocess Optimization (Figure 1). aPro is based on the business process lifecycle as described in [19], containing components for process modeling, execution, monitoring, analysis and



**Fig. 1.** Overview of the aPro architecture. Bold parts highlight focus of this work.

adaptation. During *Modeling* a process model is created using a modeling tool and stored in a process repository. Then during *Execution* the process is either executed by a process engine or by a collaboration of (legacy) application systems, which are not process-oriented. During execution, measurement of relevant metrics has to be performed, e.g. synchronous or asynchronous. In any case, a measurement is compiled into a call to a *monitoring* web service, which transfers the data to a Complex Event Processing (CEP) [12] Engine [1]. It correlates the measurements of a process instance, calculates key performance indicators (KPIs) and checks goal fulfillment. The CEP Engine then provides the processed data to the next steps: Real-time data is displayed on a dashboard and long-term data is stored in a data warehouse. During *analysis* data mining is used to find deficits and identify possible adjustments of the process. These adjustments are then used to *adapt* the process using for example ABIS [20], a tool for adaptive business processes.

In this work, we focus on the steps modeling and monitoring, though other components already exist [10]. For defining process goals, we extend BPMN 2.0 [15] with modeling elements for metrics, KPIs and goals. Multiple components are involved in aPro, and each of them needs to be configured and adapted independently. As this would result in prohibitive effort, we generate all configuration files from the defined goals, KPIs, metrics and process model. ProGoalML, the *Process Goal Markup Language*, serves as an intermediary between the different files as shown in Figure 2. We focus on the highlighted parts, describing modeling, creation of a ProGoalML file and automatic generation of measurement and result schemata as well as CEP rules, thus encompassing all steps necessary to perform basic process monitoring. This paper is structured as follows. We first describe the modeling steps and give a motivational example in Section 2. In Section 2.2 we explain the structure of ProGoalML. Section 3 describes the measuring of metrics



and the creation of CEP rules and result schemata. Section 4 describes a prototypical implementation as well as evaluation. In Section 5 we examine related work. Section 6 gives a conclusion and outlines future work.

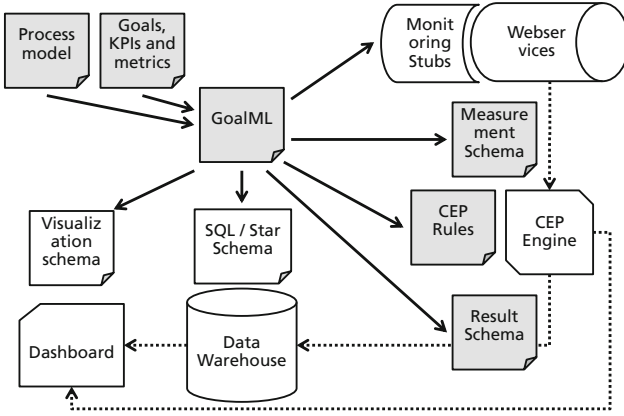


Fig. 2. Overview of documents created with ProGoalML. Dotted lines indicate flow of monitoring data between components. Grey parts highlight focus of this work.

## 2 Modeling Goals, KPIs and Metrics Using ProGoalML

In order to define a goal model and create a ProGoalML document, aPro introduces additional modeling elements used in conjunction with a BPMN diagram (see Figure 3).

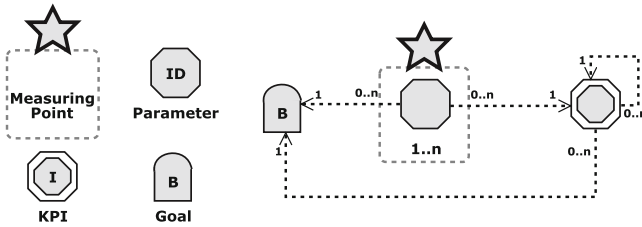


Fig. 3. ProGoalML modeling elements

To measure metrics during execution, *measuring points* are used. A *measuring point* can be attached to any BPMN element, at which a measurement is to be taken. It contains one or more *Parameters* which are to be measured. A *Parameter* has a name and a primitive data type. A special case is the ID type, which is used to correlate measurements of a process instance. Based on Parameters, KPIs can be calculated. A KPI defines a function which references Parameters or other KPIs. A *goal* is similar to a KPI as it is defined by a function as well. However, this function returns a Boolean value indicating whether the goal has been fulfilled or not. Abbreviations are used to indicate which type a measurement, goal or KPI has.

We define a *Parameter*  $p$  as a tuple  $p \in P = (n, t, E_p)$

where  $n$  is the name of the *Parameter* and the type  $t \in T$ , with  $T = \{\text{Boolean, Enumeration, Integer, Double, String, ID, Long}\}$  as the set of types. If  $t = \text{Enumeration}$ , then  $E_p$  is the set of possible enumeration values, else  $E_p = \emptyset$ .

We define a *measuring point*  $m$  as a tuple

$$m \in M = (e, P_m)$$

where  $e$  is the BPMN element the measuring point is attached to and  $P_m$  is the set of Parameters belonging  $m$ . As a parameter may only belong to a single *measuring point*, given  $P$  as the set of all parameters the following must hold

$$\forall m_1, m_2 \in M : P_{m_1}, P_{m_2} \subset P \wedge m_1 \neq m_2 \rightarrow P_{m_1} \cap P_{m_2} = \emptyset$$

We further define a KPI  $k$  as a tuple

$$k \in K = (n, f, V_k)$$

where  $n$  is the Name of the KPI,  $f$  is the function to calculate the KPI and  $V_k \subset P \cup K$  is the set of input variables. Similarly, a goal is defined as

$$g \in G = (n, f, V_g)$$

where  $n$  is the name of the goal,  $f$  is a function returning a Boolean and  $V_g \subset P \cup K$  is the set of input variables. For the purpose of KPI calculation it is necessary that there are no cyclic input variables, as otherwise no order of calculation may be found. To ensure this, the following must hold:

$$\forall k \in K : \forall v \in V_k : \exists v_0, v_1, \dots, v_n \in K : (v_0 = v_n = v \wedge \forall i \in [0, n) : v_i \in V_{v_{i+1}})$$

## 2.1 Motivational Example

Figure 4 shows a simplified claim handling process of a car insurance company which checks if a claim is justified. In the first step the claim is entered by an employee in a claims management system containing among others the stipulated amount. The second step is performed by an expert system and calculates a reference amount for the claim based on the address given by checking repair shop and rental car prices. In the last step a report is generated in the claims management system and a decision about the claim is made. Either the claim is accepted, accepted with a reduced amount, rejected or an error occurred.

In order to monitor this process, *measuring points* are defined at the activities. The first measuring point at *Enter Claim* contains two *IDs*, *ClaimID* and *Address*, as well as two other parameters: *Timestamp* of type long indicating the time the claim was entered and *Amount* of type double, the amount stipulated by the claim. The second measuring point at *Calculate Reference Amount* contains an ID named *Address*, which is the same address as in the first step and another parameter *Amount* of type double, the reference amount calculated by the expert system. The third measuring point at *Decide Claim* contains the same *claimID* as the first measuring point and two other parameters: A *Timestamp* indicating when the claim was decided and the *Result* of the Decision as an Enumeration containing the values *ACCEPTED, PARTIALLY\_ACCEPTED,*

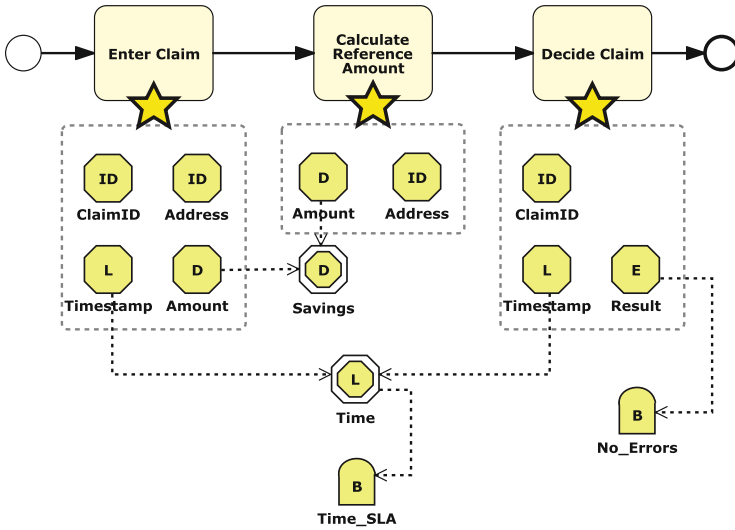


Fig. 4. Example process annotated with measuring points, KPIs and goals

REJECTED and ERROR. Measuring points are named after the elements they are attached to in order to uniquely define their parameters and measurements.

Based on these three measuring points two KPIs and two goals are calculated. The first KPI is the process execution *time* measured in seconds starting from the moment the claim has been entered, calculated from both timestamp values using the function

$$\text{Time} := (\text{Decide\_Claim.Timestamp} - \text{Enter\_Claim.Timestamp})/1000.0$$

Based on this KPI, a goal named *Time\_SLA* is defined, mandating the execution time to be below 60 seconds:

$$\text{Time\_SLA} := \text{Time} < 60.0$$

The second KPI, *Savings* achieved by the expert system is calculated using the function

$$\text{Savings} := \text{Enter\_Claim.Amount} - \text{Calculate\_Reference\_Amount.Amount}$$

The second goal mandates that the process finishes without an error:

$$\text{No\_Errors} := \text{Decide\_Claim.Result} \neq \text{"ERROR"}$$

Note that when parameters from different measuring points are used in a function, the measurements have to be *correlated* with each other using IDs.

## 2.2 ProGoalML

After *measuring points*, *KPIs* and *goals* have been modeled, a ProGoalML document has to be created to serve as an input for configuration document creation as shown in Figure 2.

Abridged ProGoalML document from motivational example (Figure 4)

```

<Progoalml version="1.0">
  <Meta>...</Meta>
  <GoalModel>
    <MeasuringPoint name="Calculate_Reference_Amount">
      <RefBpmn>Calculate_Reference_Amount</RefBpmn>
      <Parameter name="Address">
        <DataType>ID</DataType>
      </Parameter>
      <Parameter name="Amount">...</Parameter>
    </MeasuringPoint>
    <MeasuringPoint name="Decide_Claim">...</MeasuringPoint>
    <MeasuringPoint name="Enter_Claim">...</MeasuringPoint>
    <KeyPerformanceIndicator name="Savings">
      <Formula>
        Enter_Claim.Amount - Calculate_Reference_Amount.Amount
      </Formula>
      <RefParameter>
        <ParameterName>Amount</ParameterName>
        <MeasuringPointName>Enter_Claim</MeasuringPointName>
      </RefParameter>
      ...
      <DataType>double</DataType>
    </KeyPerformanceIndicator>
    <KeyPerformanceIndicator name="Time">...</KeyPerformanceIndicator>
    <Goal name="No_errors">
      <Formula>Result != "ERROR"</Formula>
      <RefParameter>...</RefParameter>
      <DataType>boolean</DataType>
    </Goal>
    <Goal name="Time_SLA">...</Goal>
  </GoalModel>
  <ProcessModel>...</ProcessModel>
</Progoalml>

```

A ProGoalML document consists of a goal model, a process model and meta-data like title and creation date. The *process model* is created by removing all aPro-related elements from the process and serializing the resulting standard BPMN 2.0 model. The *goal model* consists of measuring points, KPIs and goals. A *measuring point* contains its parameters as well as a reference to the BPMN element it belongs to. A parameter consists of a name, a data type and, if it is an enumeration, all possible enumeration values. KPIs consist of a data type, a formula and references to all input variables. Parameters are referenced by their name and the name of the measuring point they belong to. Similar to KPIs goals consist of a data type, a formula and references to all input variables, though the data type has to be *boolean*. However, in future work other data types may be supported to measure partial fulfillment of goals.

### 3 Measuring and Result Creation

As shown in Figure 1, monitoring data has to be gathered from *application systems*. To instrument these diverse application systems, different monitoring stubs are necessary, ranging from simple web service calls to periodically evaluating local log files. For each measuring point a separate monitoring stub may be created befitting the particular system performing the corresponding step.

In any case, whenever a measurement occurs, the monitoring stub calls its corresponding monitoring web service, transmitting all measured parameters. The monitoring web service then transfers the measurement to the CEP engine as an *event*. This way, CEP engine and monitoring stub are decoupled, resulting in simplified and engine-independent stub code.

*Measurement schema for Calculate\_Reference\_Amount (see Figure 4)*

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="Calculate_Reference_Amount"
    type="Calculate_Reference_Amount"/>
  <xs:complexType name="Calculate_Reference_Amount">
    <xs:sequence>
      <xs:element name="refBPMN" type="xs:string"/>
      <xs:element name="Address" type="xs:id"/>
      <xs:element name="Amount" type="xs:double"/>
    </xs:sequence>
  </xs:complexType>
</xs:schema>
```

For each measuring point an XML schema is generated, describing the structure of a measurement, called a *measurement schema* (see Figure 2). It defines the interface between monitoring stub, web service and CEP engine. In our example (see Figure 4) three schemata are generated in total.

In order to get monitoring results the CEP engine will gather measurements and calculate KPIs and goals. As KPIs and goals may be calculated from Parameters belonging to multiple *measuring points*, it is necessary to correlate all measurements from a process instance in order to obtain a matching set of input variables. For example in Figure 4 the *Savings* are calculated from two separate *measuring points*. For correlation Parameters of the type  $t = ID$  (i.e. IDs) are used. IDs with the same name are considered to have identical values in a single process instance, thus can be used to correlate their *measuring points* with each other, finding measurements which belong to the same process instance. Utilizing transitivity, measuring points with different IDs may be correlated as well. Consider our motivational example (see Figure 4). Two kinds of ID are used, *CaseID* and *Address*. The measuring point at *EnterClaim* contains both IDs, so it can be correlated to both other measuring points, which contain one of the IDs each. Thus, all three measurements which occur in a process instance can be correlated and KPIs spanning multiple measurements can be calculated.

We define the coverage class of a measuring point  $m_x$  as follows

$$C_{m_x} = \{m \in M \mid \exists m_o, \dots, m_n \in M : m_n = m_x \wedge \forall i \in [0, n) : \exists p_1 \in P_{m_i}, p_2 \in P_{m_{i+1}} : n_{p_1} = n_{p_2} \wedge t_{p_1} = t_{p_2} = ID\}$$

KPIs and goals may only use input variables belonging to one coverage class.

If a measuring point  $m_1$  is contained in the coverage class of another measuring point  $m_2$ , their coverage classes are identical:

$$m_1 \in C_{m_2} \Rightarrow m_2 \in C_{m_1} \Rightarrow C_{m_1} = C_{m_2}$$

Thus, in our example, all coverage classes are identical and contain all three measuring points.

*Algorithm to find distinct coverage classes*

```

I := {x|x ∈ P ∧ t_p = "ID"}
C := ∅
f : I ↦ C_m : f(id) := ∅
foreach (m ∈ M)
  C_m := {m}
  foreach (p ∈ P_m ∩ I)
    if (f(p) ≠ ∅)
      C := C \ {f(p)}
      C_m := C_m ∪ f(p)
    endif
  endfor
  C := C ∪ {C_m}
  foreach (p ∈ {p|p ∈ I ∧ ∃x ∈ C_m : p ∈ P_x})
    f(p) := C_m
  endfor
endfor

```

To generate result schemata and CEP rules we need to identify the set  $C$  of all distinct coverage classes using the algorithm above. Whenever it encounters a measuring point  $m$  which has an ID  $p$  already present in another coverage class  $f(p)$ , coverage classes are merged, thus ensuring every measuring point sharing an ID is in the same coverage class and only distinct coverage classes remain in  $C$ . For each coverage class a CEP rule for process instance correlation and a result schema is generated, as shown in Figure 2.

A result schema contains all goals, KPIs and parameters of a coverage class. The name of a measurement is assembled from measuring point and parameter names. IDs are named differently, as they have the same values among all measuring points and their names are globally unique.

*Abridged result schema for motivational example (compare Figure 4)*

```

<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="Handle_Claim"
    type="Handle_Claim"/>

```

```

<xs:complexType name="Handle_Claim">
  <xs:sequence>
    <xs:element name="Case_ID" type="xs:id"/>
    <xs:element name="Address" type="xs:id"/>
    <xs:element name="Enter_Claim.Timestamp" type="xs:long"/>...
    <xs:element name="Savings" type="xs:double"/>...
    <xs:element name="No_Errors" type="xs:boolean"/>...
  </xs:sequence>
</xs:complexType>
</xs:schema>

```

Similar to SQL and tables, the Esper Event Processing Language allows defining a statement on incoming events, called a rule. This rule uses the IDs to correlate all measurements from a process instance, calculates KPIs and goals and creates a result event. For each process instance, the CEP rule creates a result according to the schema. These messages may then be stored in a data warehouse, displayed on a dashboard or further aggregated within the CEP engine (future work).

*Abridged CEP rule to create results for motivational example (Figure 4)*

```

INSERT INTO Manage_Claim SELECT A.ClaimID as ClaimID, ...,
A.Amount - B.Amount as Savings, ..., C.Result != "ERROR" as No_Errors, ...,
A.Amount as Enter_Claim.Amount, ..., C.Result as Decide_Claim.Result
from pattern [every A=event(refBPMN="Enter_Claim")
-> B=event(Address=A.Address and refBPMN="Calculate_Reference_Amount")
-> C=event(ClaimID=A.ClaimID and refBPMN="Decide_Claim")]

```

## 4 Prototype and Evaluation

To evaluate ProGoalML, we created a prototype for modeling goal models, as well as generating ProGoalML files, measuring and monitoring schemata and CEP rules. The prototype is based on Oryx, a web-based tool for collaborative modeling [7]. It has been used to model the motivational example in Figure 4.

We then created a test driver generating random measurements from a given ProGoalML file. We evaluated the documents created by our prototype in multiple examples and found them to be correct. However, when placing a measuring point inside a loop, multiple measurements per process instance may occur.

Further on, we created an interactive test driver allowing tests with more realistic data. We extended the motivational example to the real-world process it represents and tested it using sample data gained by studying the real system. This resulted in correct results as well and we plan to instrument the production system in order to further evaluate GoalML.

## 5 Related Work

Goal modeling is a topic in requirements engineering [11], where goals for a (future) system are defined. Compared to ProGoalML goals are defined a priori, before the system in question exists. Goal models are then used to find goal

conflicts, identify necessary requirements and ensure requirement completeness. Goals in requirements engineering are linked and may support or contradict each other [5]. In [9] an overview of goal-oriented requirements engineering is given. Goals may be formalized using temporal logic which may be used as a basis for verification of a system implementing the requirements, e.g. by checking if a statement holds true at all times. In comparison, ProGoalML is used to define goals a posteriori, focusing on measuring their fulfillment rather than ensuring it, which is performed in the later steps of the aPro architecture. As ProGoalML checks goal fulfillment on a process instance level compared to a system level, there is no temporal dimension to goal definition. Aggregating goal fulfillment is performed in the later stages of aPro and subject for future research.

[6] defines a notation for modeling complex events in a BPMN process, similarly to the definition of measurement points in ProGoalML. We experimented using extended BPMN events for monitoring purposes with the BPMN engine *activiti*, but concluded that we need events from a broader range of systems.

In [17] Process Performance Indicators (PPIs), which parallel KPIs in ProGoalML, are defined using an ontology. They as well may be calculated hierarchically from measurements and may span a process instance or the whole process. In the latter case measurements from a process instance are aggregated. We plan to address aggregation of KPIs and goals in future work using further CEP rules. Similarly to ProGoalML, PPIs are to be used across the business life-cycle, but measurement and other steps are not described in detail. A graphical notation for PPIs is planned, but has not been implemented yet.

Similarly to creating KPIs from multiple values, [4] defines Service Level Objects in a hierarchical fashion in order to find causes of service level violations. Monitoring data is correlated using hierarchically structured event logs, a technique not applicable for aPro, as event logs may not exist.

In [3] a method for run-time validation of WS-BPEL processes is presented. The process is augmented with rules like pre- and post-conditions to use a *monitoring manager* as a proxy for service calls which polices rule compliance. In comparison to ProGoalML rules need to be written manually and separate from the process. Similarly, [2] monitors WS-BPEL processes by extending the runtime engine and transmitting events to a monitoring engine. Instance monitors for monitoring single instances and class monitors for gathering statistics across instances may be specified using *Monitoring Rules*, which, similarly to ProGoalML are then used to create Java Code. Rules for monitors are similar to CEP rules [12] and deliver numeric or Boolean values like KPIs and goals, but have to be specified manually separate from the process. Both approaches require a process engine.

[21] presents a top-down approach for modeling process performance metrics of a BPEL process. Process and PPM model are transformed to a monitoring model for use in a *Business Activity Monitoring* (BAM) tool and an event filter for a specific process engine, selecting events sent to the BAM tool. As in [2] instance and aggregate performance metrics are differentiated. Like ProGoalML modeling is initially platform-independent and then translated in platform-specific documents, but still requires a process engine.



[13] describes a model driven approach for monitoring of BPEL processes. The process is modeled across multiple abstraction levels, each containing an additional monitoring model. KPIs are derived from templates to facilitate reuse. The platform-independent monitoring model is modeled using Eclipse Modeling Framework and transformed to platform specific event and monitoring models. Similar to ProGoalML, the complexity of the underlying platform is hidden from the modeler, but only specific systems may be monitored (IBM Websphere).

In [18] a framework for non-intrusive monitoring is described. Similar to aPro monitoring is separated from execution and monitoring data is acquired by polling for events. A *monitoring policy* containing a process model, input event descriptions (comparable to monitoring schemata) and requirements to monitor (comparable to goals) is used to configure the monitoring framework. While this approach is less intrusive than aPro due to the lack of monitoring stubs, it requires events to be already generated in the execution environment. Creation of a *monitoring policy* is not automated and thus needs multiple documents to be written.

## 6 Future Work and Conclusion

Work on the prototype is ongoing. As shown in Figure 2 we plan automatic creation of a data warehouse and automatic configuration of a process dashboard already developed. Further research on CEP rule generation will be necessary, as multiple iterations are not handled now and aggregated data (e.g. averages of KPIs) will be shown on the dashboard. Additionally, we will research automatic generation or configuration of monitoring stubs and web services for data collection. Further on, we plan to support process analysis and integrate ABIS to achieve process adaptation (see Figure 1).

In this paper, we gave an overview of the business process monitoring requirements of aPro and designed a goal modeling notation. We showed how the goal model is transformed into a ProGoalML document and how this document is used to create necessary configuration documents. We detailed the creation of XML schemata for measurements and results as well as the creation of CEP rules to transform measurements to results. We implemented these concepts in a prototype and validated them using a test driver. In the future, our findings will be part of aPro, allowing business process optimization and fast setup of process monitoring without extensive technical knowledge.

## References

1. Esper - Complex Event Processing, <http://esper.codehaus.org/>
2. Barbon, F., Traverso, P., Pistore, M., Trainotti, M.: Run-Time Monitoring of Instances and Classes of Web Service Compositions. In: International Conference on Web Services, ICWS 2006, pp. 63–71 (September 2006)
3. Baresi, L., Guinea, S.: Towards Dynamic Monitoring of WS-BPEL Processes. In: Benatallah, B., Casati, F., Traverso, P. (eds.) ICSOC 2005. LNCS, vol. 3826, pp. 269–282. Springer, Heidelberg (2005)

4. Bodenstaff, L., Wombacher, A., Reichert, M., Jaeger, M.C.: Monitoring Dependencies for SLAs: The MoDe4SLA Approach. In: IEEE 5th Int'l Conference on Services Computing, pp. 21–29. IEEE Computer Society Press (July 2008)
5. Dardenne, A., van Lamsweerde, A., Fickas, S.: Goal-directed requirements acquisition. *Science of Computer Programming* 20(1-2), 3–50 (1993)
6. Decker, G., Grosskopf, A., Barros, A.: A Graphical Notation for Modeling Complex Events in Business Processes. In: 11th IEEE International Enterprise Distributed Object Computing Conference, EDOC 2007, p. 27 (October 2007)
7. Decker, G., Overdick, H., Weske, M.: Oryx – An Open Modeling Platform for the BPM Community. In: Dumas, M., Reichert, M., Shan, M.-C. (eds.) BPM 2008. LNCS, vol. 5240, pp. 382–385. Springer, Heidelberg (2008)
8. Gartner: Gartner Reveals Five Business Process Management Predictions for 2010 and Beyond, <http://www.gartner.com/it/page.jsp?id=1278415>
9. Kavakli, E., Loucopoulos, P.: Goal modeling in requirements engineering: Analysis and critique (2004)
10. Koetter, F., Weidmann, M., Schleicher, D.: Guaranteeing Soundness of Adaptive Business Processes Using ABIS. In: Abramowicz, W. (ed.) BIS 2011. LNBIP, vol. 87, pp. 74–85. Springer, Heidelberg (2011)
11. van Lamsweerde, A.: Goal-oriented requirements engineering: a guided tour. In: Proceedings of Fifth IEEE International Symposium on Requirements Engineering 2001, pp. 249–262 (2001)
12. Luckham, D.C.: The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems. Addison-Wesley, Boston (2001)
13. Momm, C., Gebhart, M., Abeck, S.: A Model-Driven Approach for Monitoring Business Performance in Web Service Compositions. In: Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services, pp. 343–350. IEEE Computer Society, Washington, DC (2009)
14. Neubauer, T.: An Empirical Study about the Status of Business Process Management. *Business Process Management Journal* 15(2), 166–183 (2009)
15. Object Management Group (OMG): Business Process Model and Notation (BPMN) Version 2.0 (2009), <http://www.omg.org/spec/BPMN/2.0/>
16. Patig, S., Casanova-Brito, V., Vögeli, B.: IT Requirements of Business Process Management in Practice – An Empirical Study. In: Hull, R., Mendling, J., Tai, S. (eds.) BPM 2010. LNCS, vol. 6336, pp. 13–28. Springer, Heidelberg (2010)
17. del-Río-Ortega, A., Resinas, M., Ruiz-Cortés, A.: Defining Process Performance Indicators: An Ontological Approach. In: Meersman, R., Dillon, T.S., Herrero, P. (eds.) OTM 2010, Part I. LNCS, vol. 6426, pp. 555–572. Springer, Heidelberg (2010)
18. Spanoudakis, G.: Non Intrusive Monitoring of Service Based Systems. *International Journal of Cooperative Information Systems* 15, 325–358 (2006)
19. Weber, B., Sadiq, S., Reichert, M.: Beyond Rigidity - Dynamic Process Lifecycle Support: A Survey on Dynamic Changes in Process-aware Information Systems. *Computer Science - Research and Development* 23(2), 47–65 (2009)
20. Weidmann, M., Koetter, F., Kintz, M., Schleicher, D., Mietzner, R.: Adaptive Business Process Modeling in the Internet of Services (ABIS). In: Internet and Web Applications and Services, ICIW (2011)
21. Wetzstein, B., Strauch, S., Leymann, F.: Measuring Performance Metrics of WS-BPEL Service Compositions. In: Fifth International Conference on Networking and Services, ICNS 2009, pp. 49–56 (April 2009)

# Automatic Derivation of Service Candidates from Business Process Model Repositories

Henrik Leopold<sup>1</sup> and Jan Mendling<sup>2</sup>

<sup>1</sup> Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany  
[henrik.leopold@wiwi.hu-berlin.de](mailto:henrik.leopold@wiwi.hu-berlin.de)

<sup>2</sup> WU Vienna, Augasse 2-6, A-1090 Vienna, Austria  
[jan.mendling@wu.ac.at](mailto:jan.mendling@wu.ac.at)

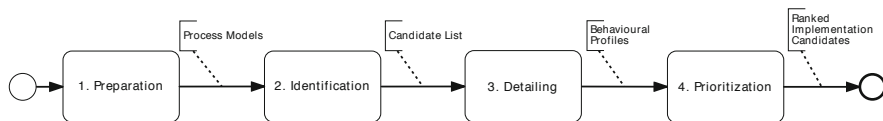
**Abstract.** Although several approaches for service identification have been defined in research and practice, there is a notable lack of automatic analysis techniques. In this paper we take the integrated approach by Kohlborn et al. as a starting point, and combine different analysis techniques in a novel way. Our contribution is an automated approach for the identification and detailing of service candidates. Its output is meant to provide a transparent basis for making decisions about which services to implement with which priority. The approach has been implemented and evaluated for an industry collection of process models.

## 1 Introduction

Services-Oriented Architecture has been discussed for roughly a decade as a concept to increase the agility of a company in providing goods and services to external partners and organizing internal operations. In this context, a service can be understood as an action that is performed by an entity on behalf of another one, such that the capability of performing this action represents an asset [1]. The focus on services is supposed to improve business and IT alignment, as it establishes principles like abstraction, autonomy and reuse [2].

A plethora of approaches to service derivation have been defined in the past. A core problem is though that many of these approaches lack methodological detail, and that none of them builds on automatic analysis techniques, cf. [2]. The problem is that a manual approach does not scale up to the size of a whole company. Indeed, several approaches recommend the manual specification of capabilities, among others based on interviews [3,4,2]. The benefits of reusing information artifacts, e.g. process models, has been recognized, among others in [5], but not in an automatic way. However, the entirety of a company can hardly be taken into account as long as models have to be manually created and inspected.

In this paper, we address the problem of manual work in the phases of service derivation. We consider the situation where an extensive set of hundreds of process models is available, which is realistic for many medium-sized and big companies [6]. Our contribution is an approach for the automatic derivation of service candidates, augmented with a set of metrics giving first clues about



**Fig. 1.** The four phases of service derivation

priorities. This approach is meant as a decision support tool for business and IT managers to quickly spot reuse potential in their company. In this way, the approach aims to speed up derivation drastically, and it can easily scale for involving large sets of process models of the whole company.

The paper is structured accordingly. Section 2 introduces the procedure of service derivation as it is summarized in related research. Section 3 introduces our approach, which builds on automatic techniques for parsing activity labels in process models. Section 4 presents the results of testing our prototypical implementation on a set of roughly 600 process models from practice. Section 5 discusses related work before Section 6 concludes the paper.

## 2 Background

This section introduces the theoretical background of service derivation. Several of the existing approaches explicitly distinguish between business services and software services. This distinction is brought forth by different perspectives. A business service can be understood as a *specific set of actions that are performed by an organization* [7], while a software service describes a part of an application system that is utilized by several entities independently [2]. The concept of a business service puts more emphasis on the economic perspective, as the software service is more related to information technology. This divide is also apparent in many of the methodological contributions on service derivation [8,9,10]. Typically, the derivation of business services tends to take more of a top-down approach, and the software service derivation is rather bottom-up. It has been shown though that both derivation types have many commonalities. For both service types, business and software services, many authors consider the analysis and evaluation of business process models to be a central step [11,12,13]. Therefore, we will focus on the reuse of process models and abstract from differences between both service types here. Accordingly, we describe service derivation as a four phase approach involving preparation, identification, detailing, and prioritization, similar to the integrated approach of [2]. Figure 1 illustrates this approach.

The derivation of services usually starts with a *preparation phase*. In this phase, an information base for the service analysis is established. This information base may include different types of business documents such as enterprise architectures, business processes or organizational structures. In this paper, we assume that a collection of process models is already available. This is a viable

assumption since big and medium-sized companies typically possess hundreds of process models [6]. The subsequent *identification phase* is concerned with identifying capabilities. In process models, these capabilities can be closely related to actions. If required, the available processes have to be further decomposed in order to arrive at a suitable level of detail. In the following *detailing phase*, the relationships and interactions between services are identified. This includes the detection of overlaps with existing services and the proper incorporation of new services into the existing SOA landscape. Finally, the *prioritization phase* is utilized to decide which services should be considered for implementation with which priority.

This four-phase process shows that the issue of scalability is hardly discussed. That is a significant problem when a service-oriented architecture is embraced as a company-wide concept. When starting from a process perspective, this means that dozens of processes have to be modeled. Often, it takes weeks to document only a single process. Even if a big number of process models already exist, it is hardly possible to inspect them manually in a systematic way. Against this background, it is striking that none of the service derivation approaches from the extensive list collected in [2] considers the potential for automation. In the following, we will define an approach that assembles analysis techniques in an innovative way towards this end.

### 3 Automatic Service Identification and Detailing

This section discusses our approach for the automatic identification and detailing of service candidates from process models. The basis for our algorithm is a set of process models  $P$  where each process model  $p$  is characterized by set of activities  $A$ . An activity  $a$  is further defined as a combination of an action  $an$  and a business object  $bo$  on which this action is performed. As an example consider the activity *Verify Invoice* which contains the action *verify* and the business object *invoice*. The union of all activity sets  $A$  from the model collection  $P$  is denoted with  $A_P$ . In order to *identify* an ordered list of service candidates  $S$  from all activities  $A_P$  we introduce a two-phase approach. In the first phase we parse all activities contained in  $A_P$  and annotate them with their according action and business object. In the second phase, we employ different strategies to identify a list of service candidates from these activities. Finally, we use behavioral profiles for a *detailing* of the service. The following sections introduce both phases in detail.

#### 3.1 Annotation of Process Model Activities

The goal of this phase is the precise annotation of activity labels with action and business object. In order to accomplish this, we employ a technique developed in prior work [14]. This technique builds on the insight that activity labels follow regular structures, so called label styles. The most frequent label styles are the verb-object and the action-noun style. The verb-object style is characterized by an imperative verb in the beginning which is followed by the business object.

Examples are *Notify Customer* or *Print Document*. In activities belonging to the action-noun style the action is not given as verb, but captured as a noun. As examples consider *Order Shipment* or *User Registration*. These examples illustrate that the structural knowledge about the label styles enables the proper extraction of action and business from activities. Accordingly, the annotation phase is further subdivided into two main steps: recognition of activity labeling style and derivation of action and business object from activity labels.

**Recognition of Label Style:** The first step is the correct recognition of the activity label style. Thereby, it is important to appropriately cope with the typical challenges of activity labels. This includes the lack of a rich sentence structure and also the zero-derivation ambiguity. The latter refers to misinterpretations due to the fact that one syntactic word can be interpreted as verb or noun (e.g. *the plan* and *to plan*). In order to adequately determine the label style, we designed a recognition algorithm which analyzes different stages of the label context [14]. As an example, consider the activity label *Plan Data Transfer* from the SAP Reference Model. By solely analyzing the label, it is not possible to decide about the label style. The activity could either instruct to *plan* a *data transfer* or to *transfer* a record of *plan data*. However, by broadening the context and considering the whole process model collection, we can learn that many other processes from the collection deal with the business object *plan data*. Accordingly, the label is classified as an action-noun label. In cases where the context of the process model collection is not sufficient, we use a word frequency list to decide whether the first word in the label is more likely to be a verb or a noun. Accordingly, it is categorized as a verb-object or action-noun label.

**Derivation of Action and Business Object:** The second step in the annotation phase is the actual derivation of action and business object from the activity label. Therefore, we make use of the structural knowledge about the label styles. Thus, we know that a verb-object label begins with an imperative verb which is followed by a business object. Accordingly, the verb-object label *Contact Customer* can be easily decomposed into the action *contact* and the business object *customer*. In the same vein, we derive action and business object from action-noun labels. As an example consider the activity *Credit Status Analysis*. Being aware that this is an action-noun label, we know that the action is given as a noun at the end of the label. By using the lexical database WordNet [15] we can derive the verb *analyze* from the nominalized action *analysis*. The business object is respectively specified with *Credit Status*.

### 3.2 Identification of Service Candidates

At this stage the action and business object from all activity labels of the considered process model collection are adequately determined. Building on this annotation information, we introduce three different approaches to identify service candidates. The following paragraphs introduce each approach in detail.

**Atomic Service Identification:** The atomic service identification strategy focuses on single activities and is based on the notion that reoccurring activities

---

**Algorithm 1.** Atomic Service Identification
 

---

```

1: List candidates = new List();
2: for each activity  $a \in A_P$  do
3:    $F_A = \text{countFrequency}(a, A_P)$ ;
4:   if  $F_A \geq 2 \wedge \text{candidates.contains}(a) = \text{false}$  then
5:      $a.\text{setFrequency}(F_A)$ ;
6:      $\text{candidates.add}(a)$ ;
7:  $\text{candidates.orderByFrequency}()$ ;

```

---

are likely to represent relevant service candidates. This approach is in line with the viewpoint of [16] that each activity in a process model can be considered as a potential service. Consequently, the frequency of a particular activity throughout the model collection determines its potential of being a suitable service candidate. In order to capture these considerations we introduce the activity frequency metric  $F_A$ , which determines the number of similar activities in a process model collection for a given activity. Thereby, the similarity between two activities is based on the congruence between their actions and business objects. Accordingly, activities following a different label style, such as *Notify Customer* and *Customer Notification*, are still considered as equal activities.

The details if the atomic service identification approach are illustrated in Algorithm 1. In order to identify services candidates for a whole process model collection, we compute  $F_A$  for each activity in the collection  $P$  (lines 2-3). If the frequency  $F_A$  of an activity is equal or greater than two and the activity has not been considered in a previous iteration, the activity is added to the candidate list (lines 4-6). After all activities have been analyzed the candidates are ordered according to their frequency (line 7). As a result, we obtain a list of candidates ordered by their potential of being suitable service candidates.

**Composite Service Identification:** The Composite Service Identification approach aims for identifying composite service candidates based on business object groups. Hence, it abstracts from single activities and focuses on activity groups having the same business object. For each business object grouping we introduce the frequency  $F_{BO}$  that determines the relevance of that group based on the occurrence of the business object among all activities of the model collection.

Algorithm 2 provides an algorithmic description for identifying composite service candidates. First, for each activity the frequency of the business object is determined (lines 2-4). In case the frequency of a considered business object is equal or greater than two and the activity - business object combination has not been stored in previous iterations, the combination is added to the group candidate map (lines 5-7). After ordering the business object groups according to their frequencies (line 8), we obtain a list of composite services candidates ordered by their relevance.

**Inheritance Hierarchy Identification:** The Inheritance Hierarchy Identification approach is based on the considerations of the Composite Service Identification strategy. However, it extends this approach by taking hierarchical



---

**Algorithm 2.** Composite Service Identification
 

---

```

1: Map groupCandidates = new Map();
2: for each activity  $a \in A_P$  do
3:    $bo = a.getBusinessObjectFromAnnotation()$ ;
4:    $F_{BO} = countFrequency(bo, A_P)$ ;
5:   if  $F_{BO} \geq 2 \wedge groupCandidates.contains(bo, a) = \text{false}$  then
6:      $bo.setFrequency(F_{BO})$ ;
7:      $groupCandidates.add(bo, a)$ ;
8:  $groupCandidates.orderByFrequency()$ ;

```

---

relationships between the business objects into account. This is motivated by the design principle of service cohesion [17] that refers to the degree of relatedness between the operations of a service. Assuming that activities with related business objects may also lead to related services, we aim for identifying business object hierarchies. In order to identify relationships between business objects, we decompose the business object terms. As an example consider the business object *purchase order*. Apparently, the word *purchase* is a specification of the main word *order* at the end. Hence, a hierarchy can be constructed by relating different parts of the business objects. For computing the relevance of such a hierarchy group we introduce the metric  $F_{IH}$  which is based on the occurrence of the main word among all business objects. The identified hierarchy groups can then be used for constructing according composite services which explicitly respect the notion of service cohesion.

Algorithm 3 illustrates the details of this approach. The basis of the hierarchy consideration are business objects which contain more than one word (line 4). If such a business object is identified, we determine the frequency of its main word among all activities (line 5). In case the frequency of the main word is equal or greater than two and no respective hierarchy tree exists, a new tree with the main word as a root node is created (lines 7-9). Afterwards, all possible business object parts are computed (lines 10-12). This is accomplished by iteratively complementing the first word of the business object until we finally obtain the original business object. Each business object part having a frequency greater or equal than two is inserted as a node on the according hierarchy level (lines 13-14). Finally, the hierarchy trees are sorted according to the frequency of the main word (line 15).

### 3.3 Detailing of Service Candidates

Service detailing refers to the definition of the structure and behaviour of a service, or a set of services. To this end, we adopt an approach for mining action patterns. Action patterns define recurring behaviour [18]. The conceptual foundation for action patterns are so-called behavioural profiles. Our approach takes a collection of process models as a starting point for deriving action patterns of a specific business object. From these patterns, we can use synthesis techniques in order to arrive at a process model that details the lifecycle of a service candidate. Therefore, this section defines the notion of a behavioural profile, explains how



**Algorithm 3.** Inheritance Hierarchy Identification

---

```

1: TreeList hierarchies = new TreeList();
2: for each activity  $a \in A_P$  do
3:    $bo = a.getBusinessObjectFromAnnotation$ ();
4:   if  $bo.getWordCount$ () > 1 then
5:      $mainWord = bo.words[bo.getWordCount$ ()];
6:      $F_{IH} = countFrequency(mainWord, A_P)$ ;
7:     if  $F_{IH} \geq 2 \wedge hierarchies.containsTree(mainWord) = \mathbf{false}$  then
8:        $mainWord.setFrequency(F_{IH})$ ;
9:        $hierarchies.createNewTree(mainWord)$ ;
10:    for  $i = 1$  to  $bo.getWordCount$ () do
11:       $term = bo.words[1] + \dots + bo.words[i]$ ;
12:       $F_{IH} = countFrequency(term, A_P)$ ;
13:      if  $F_{IH} \geq 2$  then
14:         $hierarchies.getTree(mainWord).addNode(term, i)$ ;
15:   $hierarchies.orderByFrequency$ ();

```

---

action patterns can be identified, and how a process model showing the service lifecycle can be found.

**Behavioural Profiles:** With our approach, we aim to identify service candidates that are utilized in various processes in a company. In order to detail such a service, we have to extract its behavioral constraints from different process models, and consolidate them in an appropriate way. So-called *behavioral profiles* capture such constraints on the level of pairs of activities. A behavioral profile builds on trace semantics for a process model, namely the weak order relation [19]. It contains all pairs  $(x, y)$  if there is a trace in which  $x$  occurs before  $y$ . For a process model  $p$ , we write  $x \succ_p y$ . The behavioral profile then defines a partition over the cartesian product of activities, such that a pair  $(x, y)$  is in one of the following relations:

- strict order relation  $\rightsquigarrow_p$ , if  $x \succ_p y$  and  $y \not\prec_p x$ ;
- exclusiveness relation  $+_p$ , if  $x \not\prec_p y$  and  $y \not\prec_p x$ ;
- interleaving order relation  $||_p$ , if  $x \succ_p y$  and  $y \succ_p x$ .

Based on this behavioral profile, we can define the behavioral constraints of a service candidate.

**Action Patterns:** Once we have derived the behavioral profile relations from a set of process models, we can determine the support and confidence of action patterns. This works similar to association rule mining. For each pair of activities that co-occur in one of the process models, we map them to their behavioral profile relation. Accordingly, a behavioral action pattern can be defined as a rule  $R$  with a minimum support and confidence value [18] such that:

- $R$  is a rule  $X \Rightarrow Y$ , where  $X, Y \subset V \times \{\rightsquigarrow, \rightsquigarrow^{-1}, +, ||\} \times V$ , such that  $X$  and  $Y$  are pairs of actions for which behavioral relations are specified;
- $minsup$  is the value of the required minimal support;
- $minconf$  is the value of the required minimal confidence.

Such a pattern typically captures the relationship between actions, i.e. verbs mentioned in the activity labels. We can define object-sensitive action patterns if we only consider actions of the same business object. In our context, such object-sensitive action patterns provide the basis for detailing the lifecycle of a service candidate.

**Synthesis of Service Lifecycle:** The remaining challenge is to define a process model that matches the behavioral relationships of the service candidate. A corresponding synthesis technique has been defined in [20]. The idea is to identify the consistent set of behavioral relations. From these relations, we can construct a process model. The strict order relation defines the skeleton of a corresponding process model. Activities that are not in an order relation are organized in nested XOR- or AND-blocks depending on whether they are exclusive or interleaving. The notion of profile consistency guarantees that such a nesting exists [20].

## 4 Evaluation

To demonstrate the capability of our service identification approach, we conduct an evaluation with real-world data. In particular, we designed a test collection that contains the activity labels of the SAP Reference Model. The SAP Reference Model is a collection of Event-Driven Process Chains (EPCs) and captures the business processes supported by the SAP R/3 system in its version from the year 2000 [21, pp. 145-164]. It is organized in 29 functional branches as for instance sales and accounting and contains 604 process models with in total 2433 activity labels. In the following paragraphs we present the results for each of the in subsection 3.2 introduced concepts.

### 4.1 Activity Frequency Results

For obtaining atomic service candidates we computed the metric  $F_A$  for each activity in the process model collection. As a result, we identified 464 activities with a frequency of at least two. Twelve of these activities even had a frequency of equal or greater 10. Table 1 gives an overview of the top 5 ranked results.

The results show that it is still necessary to evaluate whether an identified candidate is suitable for being established as a service. In addition, we must decide whether a candidate can be established as a business or as a software service. For instance *Process Goods Issue*, *Billing* and *Planning* are more likely to represent business services, while *Calculate Overhead* and *Difference Processing* could be automatable activities and are hence candidates for software services.

### 4.2 Business Object Frequency Results

Based on the consideration of the metric  $F_{BO}$ , we identified 378 business object groups. Thereby, each business object appeared at least twice among all activities of the model collection. In 27 of these groups the business object was used ten times or more. Table 2 shows the top 5 ranked business objects groups. In

**Table 1.** Results for Atomic Service Identification

Rank	Activity	$F_A$
1	Process Goods Issue	20
2	Calculate Overhead	17
3	Billing	13
4	Planning	13
5	Difference Processing	13

**Table 2.** Results for Composite Service Identification

Rank	Business Object	Example Actions	$F_{BO}$
1	Order	Execute, Settle, Archive, Release, Print	48
2	Time Sheet	Report, Permit, Process, Approve, Create	23
3	Invoice	Release, Verify, Process, Receive, Reverse	23
4	Budget	Release, Plan, Update, Allocate, Return	23
5	Posting	Perform, Release, Direct	19

addition to the rank and the value of  $F_{BO}$ , it also provides the most frequent actions which are applied on these business objects in the analyzed models.

### 4.3 Inheritance Hierarchy Results

To extract a business object hierarchy, we derived the different parts for each business object and computed their frequencies. In this way, for instance the business object *Service Product Order* was first reduced to *Product Order* and then to *Order*. Whenever a part term was identified twice, a new node in the business object hierarchy was created. Taking the given example, only a new node for *Order* is introduced as the term *Product Order* only appeared once among all activities of the model collection. By computing the metric  $F_{IH}$  for each main word, we obtain a ranked list of business object hierarchies.

In total, we identified 362 business object hierarchies where the main word was at least found twice among all models. In 70 hierarchies the root term was used 10 times or more. Table 3 shows the top 5 ranked business object hierarchies including the main word and three frequent example nodes.

### 4.4 Determination of Internal Service Structure

In order to determine the internal structure of a composite service, we computed the behavioural profile for the comprised activities. Table 4 shows the behavioural profile for the composite service *order*. This profile illustrates that there exists a well-defined order in which the activities must be executed. Apart from the activity *print*, which can be performed at any time after the activity *process*, the order is strict. The synthesis of this profile yields the process model depicted in Figure 2.

**Table 3.** Results for Composite Service Identification

Rank	Main Word	Example Nodes	$F_{IH}$
1	Order	Business Order, Sales Order, Service Order	112
2	Data	Plan Data, Transaction Data, Time Sheet Data	51
3	Cost	Plan Cost, Process Cost, Shipment Cost	46
4	Request	Payment Request, Recruitment Request	30
5	Document	Billing Doc., Customer Doc., Customer Doc.	29

**Table 4.** The behavioural profile for the composite service *order*

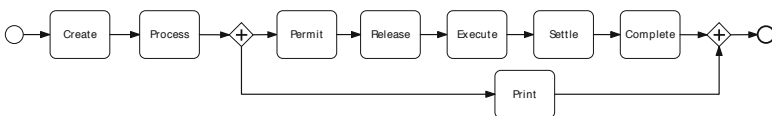
	<i>create</i>	<i>process</i>	<i>print</i>	<i>release</i>	<i>execute</i>	<i>permit</i>	<i>settle</i>	<i>complete</i>
<i>create</i>		↔	↔	↔	↔	↔	↔	↔
<i>process</i>			↔	↔	↔	↔	↔	↔
<i>print</i>								
<i>release</i>					↔	↔ <sup>-1</sup>	↔	↔
<i>execute</i>						↔ <sup>-1</sup>	↔	↔
<i>permit</i>							↔	↔
<i>settle</i>								↔

## 5 Related Work

The work presented in this paper relates to approaches for service identification and the application of natural language processing for process models.

Service identification has been considered for business services and software services [2]. The identification of business services is mainly discussed by papers from practice [22,23]. Those typically build on the analysis of business entities and components. By contrast, the derivation of software services is widely discussed in research. Some authors propose a bottom-up approach by analyzing existing legacy systems and building on their functionality [24,25]. Others suggest a top-down strategy [9,26]. However, most approaches propose to strike a balance between the two extremes. As a result, plenty of software service identification approaches include the analysis and evaluation of business process models [11,12,13].

Techniques for natural language processing have been applied on process models in various contexts. One important application scenario is the assurance of linguistic quality aspects in process models. With this intention NLP tools were employed for identifying semantic errors in activity labels [27] and for constructing

**Fig. 2.** The process model for the composite service *order*

a glossary from process model collections [28]. NLP tools were also used to refactor whole process model collections in order to ensure the understandability of the comprised activity labels [14]. As the latter requires the decomposition of the activity label into its components, this technique also constituted the basis for the approach presented in this paper. Other prominent applications of NLP techniques are the identification of similarities between process models [29,30] and the derivation of process models from natural language texts [31,32,33].

## 6 Conclusion

In this paper we presented an approach for the automatic derivation of service candidates from process models. We built on analysis techniques for process models and proposed three different techniques for deriving service candidates. We tested our approach on a process model collection from practice including 600 EPCs with 2433 activities. The evaluation illustrates the capability of our algorithm to provide useful information for service derivation. Our technique does not only enable companies to take an extensive number of process models into account, but also to efficiently analyze them. Considering the results it is important to emphasize that our technique cannot completely automate the service derivation procedure, as the final decision making remains a human task.

In future research, we plan to extend our technique by incorporating lexical relationships such as synonymy and homonymy. Further, we aim to apply our technique in the context of an industrial case study. In this way, we aim for determining the applicability and the significance of each of the identification strategies. In response to the findings, the proposed approach could then be adapted to the specific needs from practice.

## References

1. O'Sullivan, J., Edmond, D., ter Hofstede, A.H.M.: What's in a service? *Distributed and Parallel Databases* 12(2/3), 117–133 (2002)
2. Kohlborn, T., Korthaus, A., Chan, T., Rosemann, M.: Identification and analysis of business and software services - a consolidated approach. *IEEE T. Services Computing* 2(1), 50–64 (2009)
3. Hafeez, K., Zhang, Y., Malak, N.: Determining key capabilities of a firm using analytic hierarchy process. *Int. J. of Production Economics* 76(1), 39–51 (2002)
4. Homann, U., Tobey, J.: From capabilities to services: Moving from a business architecture to an it implementation (2006)
5. Zimmermann, O., Krogdahl, P., Gee, C.: Elements of service-oriented analysis and design. *IBM developerworks* (2004)
6. Rosemann, M.: Potential Pitfalls of Process Modeling: Part A. *Business Process Management Journal* 12(2), 249–254 (2006)
7. Feuerlicht, G.: Design of service interfaces for e-business applications using data normalization techniques. *Inf. Syst. E-Business Management* 3(4), 363–376 (2005)
8. Bell, M.: *Service-oriented modeling. Service Analysis, Design and Architecture*. John Wiley and Sons, Hoboken (2008)
9. Erl, T.: *Service-Oriented Architecture: Concepts, Technology, and Design*. Prentice Hall PTR, Upper Saddle River (2005)

10. Ramollari, E., Dranidis, D., Simons, A.: A survey of service oriented development methodologies. 2nd europ. young researchers WS on service oriented comp. (2007)
11. Azevedo, L.G., Santoro, F., Baião, F., Souza, J., Revoredo, K., Pereira, V., Herlain, I.: A Method for Service Identification from Business Process Models in a SOA Approach. In: Halpin, T., Krogstie, J., Nurcan, S., Proper, E., Schmidt, R., Soffer, P., Ukor, R. (eds.) BPMDS 2009 and EMMSAD 2009. LNBP, vol. 29, pp. 99–112. Springer, Heidelberg (2009)
12. Klose, K., Knackstedt, R., Beverungen, D.: In: Identification of Services - A Stakeholder-Based Approach to SOA Development and its Application in the Area of Production Planning. University of St. Gallen (2007)
13. Erradi, A., Kulkarni, N., Maheshwari, P.: Service Design Process for Reusable Services: Financial Services Case Study. In: Krämer, B.J., Lin, K.-J., Narasimhan, P. (eds.) ICSSOC 2007. LNCS, vol. 4749, pp. 606–617. Springer, Heidelberg (2007)
14. Leopold, H., Smirnov, S., Mendling, J.: On the refactoring of activity labels in business process models. *Information Systems* (forthcoming, 2012)
15. Miller, G.: WordNet: a Lexical Database for English. *CACM* 38(11), 39–41 (1995)
16. Inaganti, S., Behara, G.K.: Service identification: BPM and SOA handshake. *BP-Trends* (2007)
17. Papazoglou, M.P., Heuvel, W.V.D.: Service-oriented design and development methodology. *Int. J. Web Eng. Technol.* 2, 412–442 (2006)
18. Smirnov, S., Weidlich, M., Mendling, J., Weske, M.: Action patterns in business process model repositories. *Computers in Industry* 63 (2012)
19. Weidlich, M., Mendling, J., Weske, M.: Efficient consistency measurement based on behavioral profiles of process models. *IEEE T. Softw. Eng.* 37(3), 410–429 (2011)
20. Smirnov, S., Weidlich, M., Mendling, J.: Business process model abstraction based on synthesis from consistent behavioural profiles. *Int. J. Coop. Inf. Sys.* 21 (2012)
21. Keller, G., Teufel, T.: SAP(R) R/3 Process Oriented Implementation: Iterative Process Prototyping. Addison-Wesley (1998)
22. IBM: Component business models (2005)
23. SAP: Enterprise service design guide (2005)
24. Sneed, H.M.: Integrating legacy software into a service oriented architecture. In: *IEEE Conference on Software Maintenance and Reengineering*, pp. 3–14 (2006)
25. Belushi, W.A., Baghdadi, Y.: An Approach to Wrap Legacy Applications into Web Services. In: *Int. Conf. on Service Systems and Service Management*, pp. 1–6 (2007)
26. Flaxer, D., Nigam, A.: Realizing business components, business operations and business services. In: *Proceedings of IEEE CEC-EAST*, pp. 328–332 (2004)
27. Gruhn, V., Laue, R.: Detecting Common Errors in Event-Driven Process Chains by Label Analysis. *Enterprise Modelling and Inf. Systems Arch.* 6(1), 3–15 (2011)
28. Peters, N., Weidlich, M.: Automatic Generation of Glossaries for Process Modelling Support. *Enterprise Modelling and Inf. Systems Architectures* 6(1), 30–46 (2011)
29. Dijkman, R., Dumas, M., van Dongen, B., Käärrik, R., Mendling, J.: Similarity of business process models: Metrics and evaluation. *Inf. Syst.* 36, 498–516 (2011)
30. Ehrig, M., Koschmider, A., Oberweis, A.: Measuring Similarity between Semantic Business Process Models. In: *APCCM 2007*, vol. 67, pp. 71–80 (2007)
31. Friedrich, F., Mendling, J., Puhmann, F.: Process Model Generation from Natural Language Text. In: Mouratidis, H., Rolland, C. (eds.) *CAiSE 2011*. LNCS, vol. 6741, pp. 482–496. Springer, Heidelberg (2011)
32. de AR Gonçalves, J.C., Santoro, F.M., Baião, F.A.: Business Process Mining from Group Stories. In: *CSCWD 2009*, pp. 161–166. *IEEE Computer Society* (2009)
33. Sinha, A., Paradkar, A.: Use Cases to Process Specifications in Business Process Modeling Notation. In: *IEEE Int. Conference on Web Services*, pp. 473–480 (2010)

# Adaptability of Service Based Workflow Models: The “Chained Execution” Architecture

Saida Boukhedouma<sup>1,2</sup>, Zaia Alimazighi<sup>1</sup>, Mourad Oussalah<sup>2</sup>, and Dalila Tamzalit<sup>2</sup>

<sup>1</sup> USTHB- FEI- Departement of Computer Science, LSI Laboratory, ISI Team  
El Alia BP n°32, Bab Ezzouar, Algiers, Algeria  
{Sboukhedouma, zalimazighi}@usthb.dz

<sup>2</sup> Nantes University, LINA Laboratory, MODAL Team  
2, Rue de la Houssinière, BP 92208, 44322 – Nantes, cedex 3- France  
{Mourad.oussalah, Dalila.tamzalit}@univ-nantes.fr

**Abstract.** Business processes are frequently subject to changes which must be supported by process models and systems implementing them. This paper deals with adaptability of Inter-Organizational Workflow (IOWF) process models based on services. It states conceptually, typical adaptations that can be operated on IOWF models obeying to the chained execution architecture. IOWF models are described through the concepts of service and orchestration function expressed using basic control flow operators. Thus, operations of adaptation turn to modification of services and transformation of orchestration functions describing the model. We particularly distinguish evolvable adaptation leading to expansion of the cooperation and/or the global functionality of the process.

**Keywords:** IOWF, Chained execution, Service, Cooperation pattern, Orchestration function, Adaptation, Evolution.

## 1 Introduction

The B2B cooperation was initially supported by concepts and tools of *Inter-Organizational workflow* (IOWF) [1] and more recently by *Service Oriented Architectures* (SOA) and web services [2]. Also, many research works have been directed towards the combination of these technologies for the development of collaborative business applications. These last implement two kinds of cooperation: *ad-hoc cooperation* appropriate for non-durable cooperation and process models not completely defined at build time; or *structured cooperation* which is suitable for durable cooperation and clearly defined process models at build time.

In our research work, we are interested in structured cooperation supported by the concept of IOWF. In [1], generic architectures of IOWF have been defined; we talk about the *capacity sharing*, the *chained execution*, the *subcontracting*, the *case transfer*, the *extended case transfer* and the *loosely coupled WF*. We consider these architectures as basis of our research work because they cover a wide range of business processes since they express the different ways in which businesses can

cooperate together. However in their initial form, these architectures were subject to criticisms because of their rigidity and the difficulty to adapt business processes to support changes. Then, our idea is to propose *cooperation patterns* based on services suitable to the basic architectures defined in [1], using a *SOA based approach*. According to constraints relative to IOWF architecture, this last can be implemented through *global orchestration* or *distributed local orchestrations* of services. *Global orchestration* means that services of different partners are orchestrated using a global WF process implemented at one site; where *distributed local orchestrations* mean that services of each partner are orchestrated by a WF process implemented locally. The goal behind the use of SOA is to obtain process models flexible enough to ease their adaptation because services are loosely coupled components and platform independent.

This paper deals with *adaptability* of IOWF process models suitable to structured cooperation; we focus particularly on adaptability relative to the control flow perspective. Also, according to various reasons of adaptation, we distinguish several types, then we talk about *perfective adaptation* [3] in case of improvement of the process in order to meet the client's requirements, we talk about *adaptive adaptation* [3] in case of new constraints to take into account and we talk about *corrective adaptation* [3] if we need to correct errors in the process model. In our case, we globally talk about *adaptation of process models*. Another reason of adaptation is the evolution of process models called *evolvable adaptation* that we perceive through two perspectives: expansion of process *functionalities* and expansion of *cooperation*; we globally talk about *evolution of process models*.

The present work focuses on the *chained execution* architecture which connects two or more WFs in *sequential* manner. The paper describes the corresponding cooperation pattern based on services, states conceptually typical adaptations that can be operated on IOWF process models and describes the transformation of the orchestration function for each kind of adaptation.

For the rest of the paper, Section 2 presents some related works and explains the motivation of our work. Section 3 synthesizes the necessary background to understand the paper. Section 4 describes the *chained execution pattern* based on services and illustrates the concept of *orchestration function*. Section 5 and 6 describe respectively the different operations of adaptation and evolution of IOWF process models. Section 7 concludes the paper and talks about future works.

## 2 Related Works and Motivation

Many research works deal with the combination of WF, SOA and web services technologies for the development of flexible business collaborative applications [4], [5], [6]. This had a great impact in promoting B2B relationships since several approaches and platforms have been proposed to support business cooperation using WF and SOA. In *structured* cooperation for example, we can cite some approaches like CoopFlow [7], CrossFlow [8], CrossWork [9], Pyros [10] and e-Flow [11].



Also, flexibility is an important propriety to be satisfied by business processes and their systems allowing them to support changes. Even if some approaches like CoopFlow, Pyros and e-Flow provide *internal adaptation* of workflows without compromising the coherence of the global process, a large number of the proposed solutions are not flexible enough because they are closely coupled with the platforms. So for any changes, they impose to re-adapt the interfaces and to newly build the structure of interaction. Moreover, WF flexibility is perceived at two complementary levels: (i) at the *system level*, the flexibility defines the ability of WFMS (WF management system) to face unexpected and erroneous situations [12], [13], [14]. (ii) at the *level of process models* that defines the ability of a process model to be adaptable, evolvable and reusable. For that, many research works have been proposed describing different techniques such as adaptation patterns [15], [16], [17], rule-based adaptation patterns [18] and constraint-based modeling [19].

The goal of this paper is to deal with *adaptability of IOWF process models* based on services especially obeying to the *chained execution* architecture. First, we introduce the concept of *cooperation pattern* that we define through two dimensions: the partitioning of the process among the partner's sites and the control of execution. Then, we express this cooperation pattern using *SOA approach* in order to deal with IOWF models easily adaptable. The use of SOA is motivated by the fact that services are loosely coupled components, easily invoked through their interfaces, business oriented and platform independent and SOA paradigm supports integration, reuse and composition of services.

### 3 Basic Definitions and Concepts

#### 3.1 Definition and Architectures of IOWF

An IOWF can be defined as a manager of activities involving two or more workflows *autonomous*, possibly *heterogeneous* and *interoperable* in order to achieve a common business goal [1].

In [1], generic architectures of IOWF have been defined to support structured cooperation; we talk about the *capacity sharing*, the *chained execution*, the *subcontracting*, the *case transfer*, the *extended case transfer* and the *loosely coupled WF*. These architectures are characterized according to two main dimensions: the *partitioning of the process* and the *control of execution*. The partitioning of the process defines the way in which the process fragments of IOWF are distributed among the partner's sites (*process partitioning*) and the location of process instances at runtime (*instance partitioning*). The second dimension which is the *control of execution* defines the manner in which the execution of process instances is managed by the systems of partners. The control is *centralized* if the execution of process instances is delegated to one system that also manages all interactions between the systems of partners. The control is *decentralized* if the execution of instances is distributed among the systems of all partners and each system manages itself its interactions with other systems. We say that a control is *hierarchized* if each system manages its own WF and there is one principal system that controls interactions with

one or more secondary systems. In some cases, the control can be a *mixture* of previous modes. The *chained execution* architecture supports a model of cooperation that connects two or more business partners, each of which implements its own WF process. Workflows implied in the cooperation are executed in *sequential*. The results of execution of  $WF_i$  are input data of  $WF_{i+1}$ . In this architecture, we have *process partitioning* since each partner implements a fragment of the global WF and *instance partitioning* because at each moment a process instance is at one location; the control of execution is *decentralized*.

### 3.2 IOWF Meta-model, Adaptability and Evolutivity

An IOWF process model is defined by a set of WF fragments and a *cooperation pattern* (see Fig. 1). The cooperation pattern defines a specific architecture; it links two or more WF through a set of *interaction points*. Each WF is attached to a *partner*, manipulates *data* and is submitted to *condition* of control flow. A cooperation pattern is defined through the two dimensions of IOWF: the *partitioning of the process* and the *control of execution*. Through the concepts of the meta-model, the IOWF model covers four main axes: *process* (concepts of IOWF, WF, condition and cooperation pattern), *organization* (concept of partner), *data* and *interaction* (concept of interaction point). Consequently, we can affirm that the constraints of flexibility in IOWF model are not limited to one axis, but cover all axes that define it. However, the flexibility is mainly reflected in the *process* and *interaction* axes although it involves and also draws on other levels – data and organization.

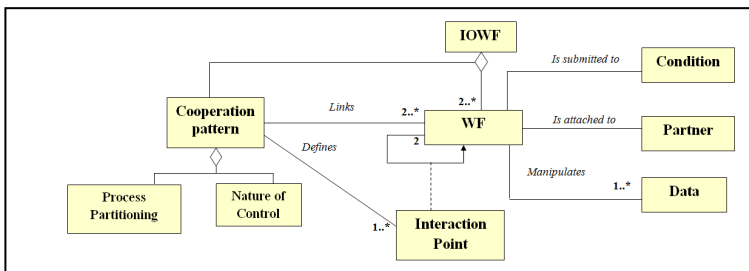


Fig. 1. Generic meta-model of IOWF

**IOWF adaptability:** An IOWF model is *adaptable* if one or more of the entities -WF, conditions, data and interaction points - composing it can be modified without affecting the global functionality and the cooperation (circle of partners).

**IOWF evolutivity:** An IOWF model is *evolvable* if it allows expansion of the global functionality or expansion of *cooperation* (additional business partners and so additional WF fragments).

As already said, we focus on the *chained execution* architecture of IOWF. For that, we describe the corresponding *cooperation pattern* (called “*chained execution pattern*”) based on services in order to deal with IOWF models easily adaptable and evolvable. Then, we introduce the concept of *orchestration function*.

## 4 Cooperation Pattern and Orchestration Function

To define a cooperation pattern suitable to a specific architecture of IOWF, the question is to decide which parts of the WF process should be encapsulated within services in order to abstract them and to invoke them from outside. Specifically, *it is to encapsulate a WF process or a sub-process in a service*. In the following, we present the *chained execution* pattern.

### 4.1 The “Chained Execution” Pattern Based on Services

For the chained execution architecture, we propose to *entirely* encapsulate WF of each partner within a service that means service  $S_i$  encapsulates  $WF_i$  provided by partner  $i$ . Process instances are executed according to the *sequence* of services implemented (see Fig. 2). Thus, the first service ( $S_1$ ) of the sequence is triggered by an external event (the occurrence of a new instance), the other services ( $S_{i+1}$ ) of the sequence, each of which is triggered by the service ( $S_i$ ) that precedes it.

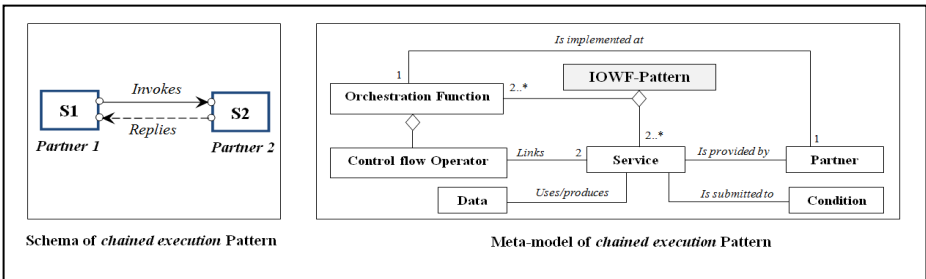


Fig. 2. Schema and meta-model of the “Chained Execution” Pattern

We can say that this architecture is implemented as *choreography* of services with *decentralized control* because services of several partners interact directly together without need to central orchestrator. Also, a reply to the service invoker can be facultative, hence the dotted arrow on the schema. The *chained execution* pattern is described through the meta-model on the right of Fig. 2, using UML notation.

At internal level, services  $S_1$  and  $S_2$  can be implemented as composite services encapsulating WFs of partner1 and partner2; it means that each internal activity of  $WF_i$  is implemented as a *local service*  $S_{ij}$ . Then, we propose implementation of a local orchestration function at each partner where maintaining a *decentralized control* of execution in the IOWF (see Fig. 3). The local orchestrator of partner  $i$  has to receive *input data* (through a service  $S_{ini}$ ) from another orchestrator to *invoke its local composite service* ( $S_i$ ) with this input data and then to invoke service  $S$  of the next partner by sending *results* (output) of its local service through service  $S_{outi}$ ; this is implemented at each partner of the IOWF.

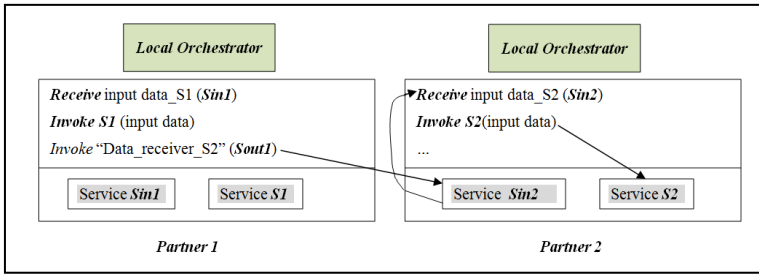


Fig. 3. Illustration of local orchestrators

### 4.2 Orchestration Function and Control Flow

On the meta-model of Fig. 2, the concept of *orchestration function* describes the control flow between services composing the WF. The orchestration function is expressed using a combination of basic control flow operators. On Fig.4, we introduce these basic operators and we express them using a general notation independently from any language or platform.

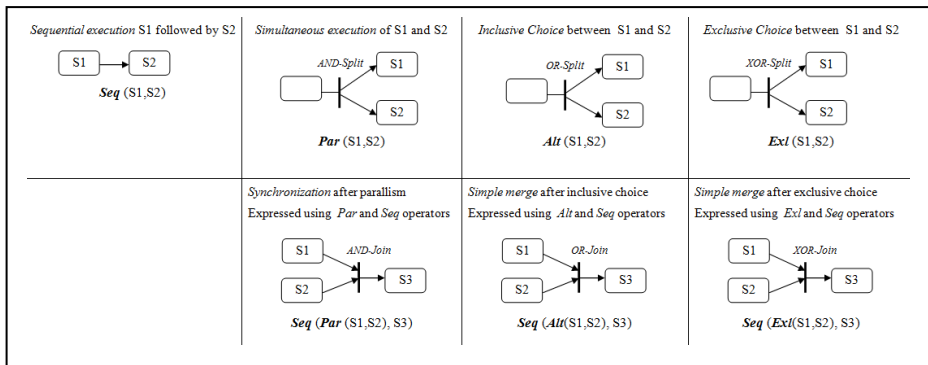


Fig. 4. Basic operators of control flow

*Remark.* To describe multi-choice – respectively multi-parallel - (more than two edges), we can decompose on several simple choices – respectively several simple parallel blocs. For example,  $Alt(S1, S2, S3)$  is expressed as  $Alt(Alt(S1, S2), S3)$  or  $Alt(S1, Alt(S2, S3))$ .

Fig. 5 bellow illustrates the concept of orchestration function using our notation; we give an example of IOWF obeying to the *chained execution pattern*. The process schema describes an IOWF implying two partners, partner 1 and partner 2 implementing their WFs as services  $S1$  and  $S2$  respectively. Partner 1 provides his WF composed by *internal* services  $S11, S12, S13, S14, S15$  and partner 2 provides his WF composed by internal services  $S21, S22$  and  $S23$ ; in this case, the service *Sout1* corresponds to invocation of  $S2$  from  $S1$ . For more readability and less complexity of the orchestration function, we can structure the process fragments into blocs  $Bij$  of

sequential, parallel or alternative services. In hierarchical manner, a bloc can be expressed using other blocs. The orchestration function can be represented by a binary tree with two types of nodes: operators and services.

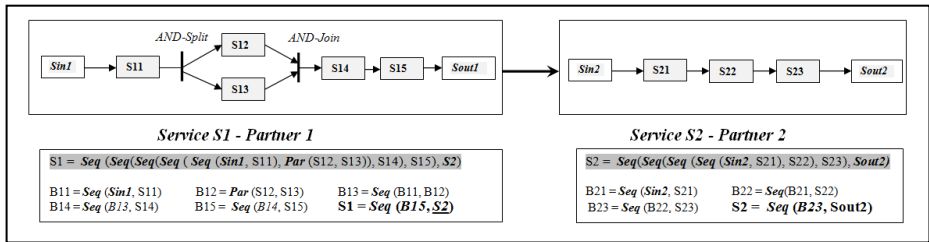


Fig. 5. Illustration of orchestration function

### 4.3 Formal Definition of IOWF

An IOWF is defined by a pair  $\langle S, F \rangle$  where S is a set of services  $S_i$  (or  $S_{ij}$  for internal level) and F is a set of orchestration functions  $f$ .  $f(S_{i1}, S_{i2}, \dots, S_{in}) = S_{i1} \text{ op1 } S_{i2} \dots \text{ opn-1 } S_{in} = S_i$  op1, ... opn-1 are operators of control flow.

For the “chained execution” pattern, an orchestration function  $f$  implemented at partner  $i$  associates a composite service  $S_i$  to a set of internal services  $S_{ij}$ . Interactions between services of the same partner define internal interactions and interactions between services of different partners define external ones.

## 5 Adaptability of IOWF Models

According to the previous definition, adaptation of process models turns to modifications of the entities composing it that means *services* or *orchestration functions*. A modification of a service can be adding, removing, replacing, merging of two services and decomposing a service into a bloc of two services expressing sequential, parallel or alternative execution. Adaptation of a service usually induces modification on the *orchestration function* using it or a modification of closely attached attributes like *condition* or *data* (see Fig. 2). Also, other operations of adaptation can affect only the control flow in the process that means the *orchestration function* while maintaining all services composing the process.

### 5.1 Adding, Removing and Substituting of Services

For *adding* or *removing* of services, it is to distinguish adding or removing of a service on *one edge* composed by sequential services or in a bloc composed by *two edges* expressing parallel or alternative execution. The part on the top of Fig. 6 describes the basic operations of *adding* of services illustrated by generic schemas, the corresponding *orchestration functions* and the sequence of operations done in order to obtain the new orchestration function from the initial one. Let’s notice that the adding of service in a bloc of exclusive choice or parallel execution is not represented in the figure because it is done in the same manner as inclusive choice.

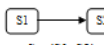
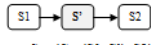
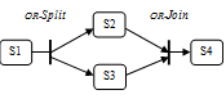
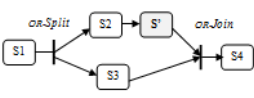
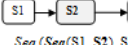
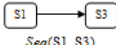




Operation of adaptation	Schema before adaptation	Schema after adaptation	Description of operations done
Add into <i>sequence</i>	 $Seq(S1, S2)$	 $Seq(Seq(S1, S'), S2)$	Create a new bloc $B = Seq(S1, S')$ in <i>sequence</i> with $S2$
Add on one edge of <i>inclusive choice</i>	 $Seq(Seq(S1, Alt(S2, S3)), S4)$	 $Seq(Seq(S1, Alt(Seq(S2, S'), S3)), S4)$	Add a service $S'$ after $S2$ to create a bloc $B = Seq(S2, S')$ in <i>alternative</i> with $S3$
Remove from <i>sequence</i>	 $Seq(Seq(S1, S2), S3)$	 $Seq(S1, S3)$	Remove the second operator <i>Seq</i> and service $S2$
Remove from one edge <u>with several services</u> of <i>inclusive choice</i>	 $Seq(Seq(S1, Alt(Seq(S2, S3), S4)), S5)$	 $Seq(Seq(S1, Alt(S3, S4)), S5)$	Remove $S2$ from a <i>sequential bloc</i> containing it
Remove from one edge <u>with single service</u> of <i>inclusive choice</i>	 $Seq(Seq(S1, Alt(S2, S3)), S4)$	 $Seq(Seq(S1, S3), S4)$	Remove $S2$ and the operator <i>Alt</i>

Fig. 6. Adding and Removing of a service

The reverse operation of adding is the *removing* of services. It is also to distinguish the removing of a service from *one edge* composed by sequential services or from a bloc composed by *two edges* according to parallel or alternative execution. Fig. 6 (the part on the bottom) shows typical operations of removing of services (service  $S2$  for example). For non sequential bloc, we only describe the removing from alternative bloc expressing inclusive choice; the same scenario is applied for exclusive choice or parallel execution. Let's notice that two configurations are possible when removing a service  $S$  from a bloc with two edges: (i) service  $S$  is in sequence with other services, (ii) service  $S$  is alone on the edge; this results on two different scenarios for operations done like shown on Fig. 6.

Another basic operation of adaptability concerns the substitution (*replacing*) of services. This is typically a *removing* of service to replace followed by an *adding* of the new service. Particularly, the *replacing* of an interactional service *Sini* or *Souti* by another is done to adapt the interface of a service  $S_i$  implied in the IOWF.

## 5.2 Fusion and Decomposition of Services

The operation of *fusion* can concern two services related by a sequence, an inclusive choice, an exclusive choice or a parallel execution, in order to simplify the process model and to abstract several services into one. The part on the top of Fig.7 describes

these basic operations, the set of operations done and the corresponding orchestration functions modified after each operation for merging  $S_2$ ,  $S_3$  in a single service  $S'$ . We can see on Fig. 7, that since services to merge are in the same bloc, they become easier to remove and to replace because the bloc  $Alt(S_2, S_3)$ ,  $Par(S_2, S_3)$  or  $Exl(S_2, S_3)$  is considered as a single *composite service to be replaced*.

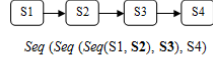
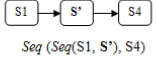
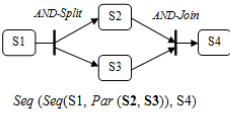
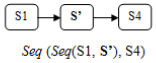


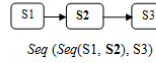
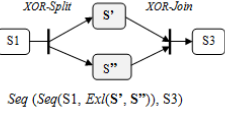
Operation of adaptation	Schema before adaptation	Schema after adaptation	Description of operations done
Fusion of <i>sequence</i>			Remove $S_2$ : $Seq(Seq(S1, S3), S4)$ Remove $S_3$ : $Seq(S1, S4)$ Add $S'$ between $S1$ and $S4$ $Seq(Seq(S1, S'), S4)$
Fusion of <i>parallel execution</i>			$S_2$ and $S_3$ are in the same bloc $B = Par(S_2, S_3)$ , we have $Seq(Seq(S1, B), S4)$ Remove bloc $B$ : $Seq(S1, S4)$ Add $S'$ between $S1$ and $S4$ $Seq(Seq(S1, S'), S4)$
Decomposition into <i>sequence</i>			Remove $S_2$ : $Seq(S1, S3)$ Add $S'$ between $S1$ and $S3$ $Seq(Seq(S1, S'), S3)$ Add $S''$ between $S'$ and $S3$
Decomposition into <i>exclusive choice</i>			Remove $S_2$ : $Seq(S1, S3)$ Add a bloc $B = Exl(S', S'')$ between $S1$ and $S3$ $Seq(Seq(S1, B), S3)$

Fig. 7. Fusion and decomposition of services

More elaborated operations of fusion concern configurations such as services to merge are not in the same bloc. For example in a model described by the function  $Seq(Seq(S1, Par(S2, S3)), S4)$ , the operation of merging  $S1$  and  $S2$  cannot be done directly since we must know if we maintain the parallelism or not; this information should be provided as additional parameter. In both cases, this must be decomposed into elementary operations of removing and adding of single services or blocs.

The reverse operation of fusion is the *decomposition* of a service to obtain a bloc of two services that can be sequential, parallel or alternative. We can see on the bottom of Fig.7 that the decomposition of a service consists to *remove* a single service ( $S_2$  for example) and to *add a bloc* composed by two services ( $S'$  and  $S''$ ) linked by sequence, alternative or parallel operator. The decomposition of services is done in order to improve the parallelism in the process (parallelization of services) or to add condition (inclusive or exclusive choice) due to new constraints or to have more control on the execution of the process (sequence of services).

### 5.3 Adapting the Orchestration Function

Another category of adaptation on IOWF models concerns modification of *orchestration function* without modifying services, this is typically a replacing of an

operator of control flow by another; we can replace for example a sequence operator (*seq*) by parallel operator (*par*) to improve the execution time of process instances, or vice versa if an execution of a service becomes dependant from another service.

When services to be restructured are in the same bloc, the operation of adaptation can be easily done by substituting operators; it is to replace the initial operator by another one in the orchestration function. For example, in the orchestration function *seq* (*seq* (*S1,S2*), *S3*), if we want to parallelize (*S1, S2*), we just replace the operator *seq* by the operator *par* to obtain the transformed function *seq* (*par*(*S1,S2*), *S3*). By contrary, if services to be restructured are not in the same bloc, operations of adaptation are less evident; for example in the orchestration function *seq* (*seq* (*seq* (*S1,S2*), *S3*), *S4*), the parallelization of (*S2,S3*) cannot be done directly but we must remove *S2* to obtain *seq* (*seq* (*S1, S3*), *S4*), then remove *S3* to obtain *seq* (*S1, S4*), and finally add a bloc *par* (*S2,S3*) between *S1* and *S4* to obtain the transformed orchestration function *seq*(*seq* (*S1,par* (*S2,S3*)), *S4*).

## 6 Evolutivity of IOWF Models

As already explained, the *evolutivity* (or evolvable adaptability) of IOWF process models is reflected at two perspectives: the *functionality* and the *cooperation* of the IOWF. Hence, an IOWF model evolves if it can be extended to additional functionalities and/or it allows expansion of cooperation to involve more partners and more external services; the two perspectives are not exclusive.

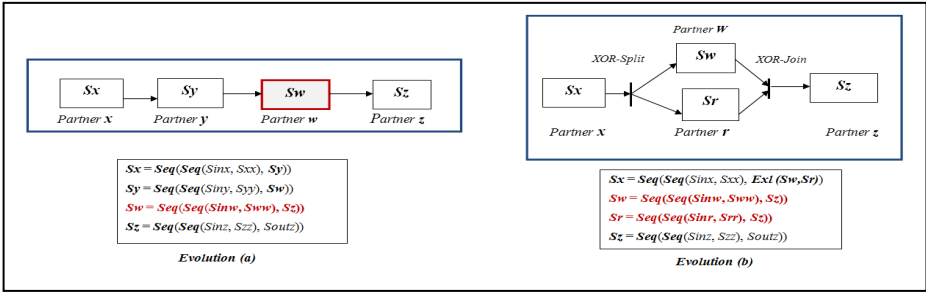
### 6.1 Expanding Functionalities

Expansion of functionalities of the IOWF can be done by adding internal services *S<sub>i</sub>* (resp. blocs) with novel functionalities into the WF of one or more partner(s) or by replacing a service (resp. bloc) by another that covers more functionalities. To do that, we can refer to operations of section 5.1, the only difference is that the injected services implement additional functionalities of the IOWF. At external level, the expansion of functionalities can be realized by replacing an external service *S<sub>i</sub>* encapsulating a WF fragment by another external service.

### 6.2 Expanding Cooperation

According to the second perspective, it is the capacity to open the IOWF to more partners. This can occur in two cases: (a) an additional external service is added to the sequence of external services composing the IOWF, in order to extend the functionality of the global process or (b) replacing an external service by a bloc of *exclusive choice* of two external services according to new constraints. Starting with an IOWF model initially composed by a sequence of three services *S<sub>x</sub>*, *S<sub>y</sub>* and *S<sub>z</sub>* provided by partners *x*, *y* and *z* respectively, Fig. 8 shows the possible configurations of evolution previously described. We assume that each service *S<sub>p</sub>* provided by partner *p* is composed by a sequence *S<sub>inp</sub>*, *S<sub>pp</sub>* (the composite business service)





**Fig. 8.** Expansion of the cooperation

and invocation of the following service in the sequence or *Soutp* for the last service in the sequence. In case of evolution (a), we have to add the orchestration function of service *w* and to ensure interaction between the pairs of services (*Sy*, *Sw*) and (*Sw*, *Sz*). In evolution (b), we have to add the orchestration functions of services *Sr* and *Sw*, to implement the exclusive choice in Service *Sx* and to ensure interaction between pairs of services (*Sx*, *Sw*), (*Sw*, *Sz*), (*Sx*, *Sr*) and (*Sr*, *Sz*). Let's notice that the *chained execution* pattern is preserved since instances are executed according to one path of sequential services (*Sx*, *Sw*, *Sz*) or (*Sx*, *Sr*, *Sz*).

## 7 Conclusion and Future Works

In this paper, we focused on the issue of adaptability of IOWF models in case of structured cooperation. We have considered process models obeying to the *chained execution* architecture defined in [1]. In order to deal with process models flexible enough, we have proposed a *cooperation pattern based on services* to implement IOWF obeying to the architecture considered in this paper. So, we have introduced the concept of *orchestration function* that is built on basic operators of control flow to orchestrate internal services composed to build a fragment of WF provided by a partner. To maintain *decentralized* control, each partner implements his orchestration function and interactional services insuring the communication with external services. We distinguish operations of evolution (evolvable adaptation) from other adaptations basis on two perspectives the *functionality* of the IOWF process and the *cooperation*; so, we talk about *evolutivity* if the functionality of the IOWF is expanded and/or the cooperation is expanded. The operations of adaptation and evolution of process models are described at a conceptual level showing the transformation of orchestration functions for each type of adaptation or evolution. Also, with the proposed approach, we can deal with reusability (well supported by SOA) of IOWF process models which is another aspect of flexibility allowing the combination of several IOWF obeying to the same or different architectures, in order to build more complex business processes based on existing ones.

We are currently working to implement these operations of adaptation and evolution as adaptation patterns by translating them to a specific language of business process definition. Furthermore, we must provide mechanisms to check the correctness of models after adaptation.

## References

1. Aalst, W.V.D.: Process oriented architectures for electronic commerce and interorganizational workflow. *Journal of Information Systems* 24(9) (1999)
2. Papazoglou, M.P., Heuvel, W.J.V.D.: Service Oriented Architectures: approaches, technologies and research issues. *The VLDB Journal* 16, 389–415 (2007)
3. Bastide, G.: SCORPIO - An Approach for Structural adaptation of software components: application to ubiquitous environments. Phd Thesis, University of Nantes (2007)
4. Chen, M., Zhang, D., Zhou, L.: Empowering collaborative commerce with web services enabled business process management system. *Decision Support System* (2005)
5. Leymann, F., Roller, D., Schmidt, M.-T.: Web Services and Business Process Management. *IBM Systems, Journal* 41(2) (2002)
6. Gorton, S., Montangero, C., Reiff-Marganiec, S., Semini, L.: StPowla: SOA, Policies and Workflows. In: Di Nitto, E., Ripeanu, M. (eds.) *ICSOC 2007*. LNCS, vol. 4907, pp. 351–362. Springer, Heidelberg (2009)
7. Chebbi, I.: CoopFlow - an approach for ascendant cooperation of workflows in virtual enterprises. Phd Thesis, National Institute of Telecom, France (2007)
8. Grefen, P., Aberer, K., Hoffer, Y., Ludwig, H.: Crossflow: Cross-organizational workflow management for service outsourcing in dynamic virtual enterprises. *IEEE Data Engineering Bulletin* 24(1), 52–57 (2001)
9. Mehandjiev, N., Stalker, I., Fessel, K., Weichhart, G.: Interoperability contributions of Crosswork. Invited short paper to Proceedings of INTEROP-ESA 2005 Conference, Geneva. Springer (February 2005)
10. Belhajjame, K., Vargas-Solar, G., Collet, C.: Pyros - an environment for building and orchestrating open services. In: Proceedings of the 2005 IEEE International Conference on Services Computing, pp. 155–164. IEEE Computer Society, Washington, DC (2005)
11. Casati, F., Shan, M.: Dynamic and adaptive composition of e-services. *Information Systems* 26(3), 143–163 (2001)
12. Sadiq, S.W., Orłowska, M.E.: On capturing Exceptions in workflow process models. In: Proceedings of ER 2001 (2001)
13. Meng, J., Su, S.Y.W., Lam, H., Helal, A., Xian, J., Liu, X., Yang, S.: DynaFlow - a dynamic inter-organisational workflow management system. *Int. Journal of Business Process Integration and Management* 1(2), 101–115 (2006)
14. Muller, R., Greiner, U., Rahm, E.: AGENT-WORK: a workflow system supporting rule-based workflow adaptation. *Journal of Data and Knowledge Engineering* 51(2), 223–256 (2004)
15. He, Q., Yan, J., Jin, H., Yang, Y.: Adaptation of Web Service Composition Based on Workflow Patterns. In: Bouguettaya, A., Krueger, I., Margaria, T. (eds.) *ICSOC 2008*. LNCS, vol. 5364, pp. 22–37. Springer, Heidelberg (2008)
16. Döhning, M., Zimmermann, B., Karg, L.: Flexible Workflows at Design- and Runtime Using BPMN2 Adaptation Patterns. In: Abramowicz, W. (ed.) *BIS 2011*. LNBIP, vol. 87, pp. 25–36. Springer, Heidelberg (2011)
17. Weber, B., Reichert, M., Rinderle-Ma, S.: Change patterns and change support features-Enhancing flexibility in process-aware information systems. *Journal of Data & Knowledge Engineering* 66, 438–466 (2008)
18. Döhning, M., Zimmermann, B., Godehardt, E.: Extended workflow flexibility using rule-based adaptation patterns with eventing semantics. In: Proc. of INFORMATIK 2010, pp. 216–226 (2010)
19. Pesic, M., Schonenberg, M.H., Sidorova, N., van der Aalst, W.M.P.: Constraint-Based Workflow Models: Change Made Easy. In: Meersman, R., Tari, Z. (eds.) *OTM 2007, Part I*. LNCS, vol. 4803, pp. 77–94. Springer, Heidelberg (2007)

# A Conceptual Model for Assessing the Benefits of Software as a Service from Different Perspectives

Anisah Herdiyanti Prabowo<sup>1,2</sup>, Marijn Janssen<sup>1</sup>, and Joseph Barjis<sup>1</sup>

<sup>1</sup> Delft University of Technology, Jaffalaan 5, 2628 BX Delft, the Netherlands  
{m.f.w.h.a.janssen,J.Barjis}@tudelft.nl

<sup>2</sup> Institut Teknologi Sepuluh Nopember, Sukolilo, 60111, Surabaya, Indonesia  
anisah@its-sby.edu

**Abstract.** Software as a Service (SaaS) has been introduced as an innovative way of software provisioning through which Web-based application is delivered as a service. Yet, SaaS model is still immature in concept and little is known about the perceptions of users and providers when understanding benefits of SaaS. For this reason, the paper explores SaaS benefits for Small and Medium Sized Enterprises (SMEs) from the perspective of providers, users during the SaaS life cycle. A conceptual model is developed and tested in a case study. Users and provider have different perceptions on the benefits of SaaS and the benefits are dependent on the maturity. Our findings show that the conceptual model can help to overcome differences in opinion about SaaS benefits at different phases of SaaS life cycle. Promised, perceived, and actual benefits might be different. The conceptual model ought to provide better structural definitions of SaaS benefits although we suggest further research to enrich the model structure by incorporating different aspects of decisions upon the benefits.

**Keywords:** Software as a Service (SaaS), Small and Medium-sized Enterprises (SMEs), Benefits, Maturity, Life cycle.

## 1 Introduction

In the past decade, many types of outsourcing models have been adopted by Information Technology (IT) managers, e.g. on-demand model and on-premise model. These models have been regarded as a cost-saving effort to efficiently adopt IT products [1], [2]. Yet, enterprises have seen significant benefits from adopting an on-demand model comparing to an on-premise model. These include more reliable services when adopting on-demand model since customers are assisted with continuous supports from service providers during the period of contract. For Small and Medium-sized Enterprises (SMEs), these benefits are of especially importance considering their limited resources and capabilities to manage IT-related functions.

Software as a Service (SaaS) has been seen as a promising opportunity to improve enterprise's revenues when providing or adopting the model [3], [4], [5]. SaaS is a model of delivering software as a service during period of contract. The services

include set-up and consultancies, and maintenance and upgrades of a Web-based application [6]. A SaaS provider delivers the application via a secured network to which users of SMEs can access on subscription basis. SaaS subscription is different from traditional business process outsourcing (BPO) which requires management of the entire business process when services are outsourced externally. Within SaaS model, maintenance, support and upgrades of a SaaS application are in the hand of SaaS providers during the period of contract while SMEs can focus on managing their core business process. In this way SMEs can have more abilities to achieve their main business goals [7] within relatively faster implementation time [8], [9]. On the other hand, SaaS providers would be able to reduce their substantial cost related to application deployment and thus they can offer faster application modifications to meet the demand of SMEs over a SaaS application.

Despite its potential benefits, SaaS adoption are risky. The risks include technical and economic-related risks [10], and psychosocial and strategic-related risks [11]. The bottom-line of these risks is that SMEs become highly dependent on SaaS providers for managing IT-related function during the period of contract of a SaaS application [12]. This dependency results into problems of creating switching cost which might occur at the end of the life-cycle [13]. Moreover, conflicting interests are due to happen because different opinions in realizing benefits of SaaS can occur at different phase of SaaS life cycle. For instance, different opinions in determining what variables that can determine cost-saving benefits when services related to SaaS application are selected and engaged, and how the benefits can be measured when these services are delivered. Furthermore, the benefits of shorter implementation time does not necessarily result in reducing investment costs since business requirements for a SaaS application can be different for enterprises, and the existing dissemination of IT within business process may be varied for different sizes of enterprises. For example, basic subscription of a SaaS application is mostly chosen by enterprises as a start-up engagement with SaaS providers. Although SaaS application can be utilized directly via online subscription agreement, the application features for this subscription is quite limited. In order to fully support their business process, enterprises usually need to add extensions and upgrade the application from basic subscription. This situation can create hidden costs which might be substantial considering vendor-dependency toward the application. These types of issues urge to understand the SaaS benefits from both the provider's and the SME's perspectives. In the light of addressing the possibility of having different perceptions when realizing benefits of SaaS, our research underlines the importance of understanding the benefits of SaaS during the whole life cycle of SaaS as hidden costs might appear in the future.

In this research, the benefits of SaaS are conceptualized from the perspectives which are the SaaS providers, users, and the life cycle. In the remainder of this paper, the related topics in the field of SaaS will be discussed briefly. According to the presented literature study, we present a conceptual model to structure the benefits of SaaS from the three perspectives. Furthermore, we also discuss a case study to implement the conceptual model, based on which we draw our conclusion of this paper. We also present discussion for further research at the end of paper.

## 2 Related Work

### 2.1 SaaS Software Service Delivery Model

SaaS is a software delivery model which is originated from the idea of IT outsourcing and subscription of services. The subscription scheme [14] can be categorized into *subscription based*, *usage based*, *transaction based*, *value based*, and *fixed fee scheme*. The *subscription based scheme* represents that service fees are paid according to the number of actual end-users of a SaaS application, while the *usage based scheme* means that service fees are paid depending on the number of servers which run the application or the number of concurrent users. The *transaction based scheme* calculates the number of transaction which is done through the application, while the *value based scheme* calculates service fees according to the achievement of business goals after the adoption of the application. Under the *fixed fee scheme*, a fixed service fee is charged according to period of services.

Within SaaS model, services for a SaaS application are standardized and can be accessed via Internet Protocol (IP)-based network. To maintain their software quality, SaaS providers improve the application continuously through a secured network and therefore they can expect that customers may prolong their subscription [15], [16]. For this reason, IT-related risks in the customers' side can be reduced since the role of managing SaaS application are transferred from customers to SaaS providers during period of contract. Customers also do not need to provide upfront investments when subscribing to a SaaS application [17] since SaaS providers account for a large amount of investments for infrastructures and application deployment [18]. On the other hand, multi-tenancy concept [19] allows SaaS providers to gain economic of scale since they can serve multiple customers with SaaS model.

Yet, high level of standardization of services and over cost reduction effort which are performed by SaaS providers, can lead to limited possibilities to customize SaaS applications. This situation is due to happen since software license is not owned by SMEs who adopt the application [20] and thus configuration options are often limited to application interface.

There are also studies which discussed about SaaS model and Service Oriented Architecture (SOA) approach. While SaaS is a software delivery model, SOA is a software-construction model [21]. SOA is considered as one of the enablers of SaaS [22], and one of the characteristics of SaaS [23]. Furthermore, SaaS application can also be offered as SOA services [24] which can optimize cloud application [25].

### 2.2 SaaS Relationship with SOC and SOA

The SaaS model provides potential benefits [26]. From the point of view of SaaS providers, investment toward SaaS deployment is considered beneficial since SaaS applications are relatively quicker and easier to market. SaaS providers can also align services related to SaaS applications by providing consultancies and training. From the point of view of SMEs who adopt the applications, continuous supports and reliable services can be expected since service levels are provided within Service

Level Agreements (SLAs) in a formal SaaS contract which also defines security and privacy of data [27]. Therefore, enterprises can ensure that services are available whenever they are needed, and at performance level they had expected. Moreover, IT-management burden is potentially reduced by transferring responsibility to manage IT function to SaaS providers. Any enterprise would expect to be successful in adopting SaaS within its business process while achieving benefits from adopting this software delivery model.

Meanwhile, enterprises are very interested to establish a set of goals which are able to define the success of SOA adoption [28]. These benefits are referred to as the benefits of Service Oriented Computing (SOC) [29]. They include increase of inherent interoperability, federation, vendor-diversification options, alignment of business and technology domains, ROI, organizational agility, and reduction of IT burden. SOC paradigm drives the SOA approach for service-oriented applications [30] while Software as a Service (SaaS) represents business models for delivering application as a service.

Nevertheless, the challenges such as how to build a good service representing these SOC strategic goals and benefits, which can adhere to the design principles of SOA, remain open. The design principles of SOA provide a set of good principles of service design, which include standardized service contract, service loose coupling, service abstraction, service reusability, service autonomy, service statelessness, service discoverability, and service composability. Standardized services within SaaS contract can support the principle of standardized service contract, while agile and interoperable services to accommodate business changes can promote the principle of service loose coupling, service reusability and service composability. Sufficient information about services represents the principle of service abstraction; whereas discovery mechanism to locate services should be provided in order to support the principle of service discovery. A service is provided when it is necessary to promote the principle of service stateless, while a service can be robust to multiple customers if the principle of service autonomy is applied.

### **3 Conceptual Model for SaaS Benefits from Different Perspectives**

As stated earlier in this paper, we introduce a conceptual model of SaaS adaption based on provider's, user's and life cycle's perspectives. By developing the model, we aim at providing a comprehensive definition of SaaS benefits which incorporate many perspectives. The model is structured by identifying characteristics of SaaS benefits and viewing these from the three perspectives. We start our study by understanding how each perspective may conceptualize benefits of SaaS. Then, we propose key characteristics of SaaS benefits from which indicators to assess the benefits can be drawn.

From the perspective of SaaS providers, the benefits of SaaS are conceptualized by considering the promised benefits which are published by SaaS providers in order to operationalize SLAs. From the perspective of SaaS users, the benefits of SaaS are

perpetually abstracted and are not well appraised since there is limited information about SaaS application in the users' side due to lack of understanding toward SaaS model which further may hinder the users to realize the benefits of SaaS. Therefore, the perspective of SaaS users is appreciated by assessing maturity level of SaaS which defines a certain benefits of SaaS at different level of benefit achievements.

The rest of this section will focus on the key characteristics of SaaS benefits which are conceptualized from the perspective of SaaS providers and users which can also be referred to as SaaS adopters. Then, both perspectives are viewed from the perspective of life cycle of SaaS from which three important phases when delivering a SaaS application will be discussed.

### 3.1 Provider Perspective

The benefits of SaaS which are perceived by SaaS users are mostly conditional according to the pre-defined benefits which are promised SaaS providers. The benefits which are perceived by SaaS users are then referred to as the perceived benefits and the benefits which are promised by SaaS providers are referred to as the promised benefits. The latter benefits are conceptualized by operationalizing level of services to be fulfilled by the providers which are agreed within SLAs in a formal SaaS contract. In this section, key attributes of SLAs will be discussed in order to define characteristics that can be used to indicate the promised benefits.

There are three key areas of SLAs [13], which are related to *service costs*, *service availability*, and *data security*. Service costs include basic service charge, additional service fees, escalating discounts for incremental spending; while service availability include period of services, up-time guarantees, business continuity in case of disaster, and termination of services and the related compensation. Data security is related to preserving enterprise and individual data, managing data ownership and security access. In addition to these key areas of SLAs, enterprises usually have a certain level of information criteria which are expected when business information is provided by IT solution and the related services. They are *effectiveness*, *efficiency*, *confidentiality*, *integrity*, *availability*, *compliance*, and *reliability* [31]. Failing to meet these criteria would likely result in information adequacy that might affect business performance.

In the light of combining key areas of SLAs and the aforementioned information criteria, we propose indicators of key benefits of SaaS with respect to the promised benefits which are offered by SaaS providers. They are *resource efficiency*, *process effectiveness*, *data confidentiality*, *integrated information*, *service availability*, *service reliability*, and *regulatory compliance*. From these indicators, we define three categorizations of the promised benefits, which are *technology-enabled business support benefits*, *interactive support benefits*, and *comprehensive support benefits*. The *technology-enabled business support benefits* represent efficient management of (enterprise) resources and effective management of business process via SaaS application; while the *interactive support benefits* represent available and reliable services, and high-secured access within SaaS application. The *comprehensive support benefits* represent integrated information which can support business, and regulatory compliance to laws that are related to the delivery of SaaS application.

### 3.2 User Perspective

In order to accommodate the perspective of SaaS users, a SaaS maturity model is developed in order to define to what extent benefits of SaaS can be expected by users at different level of maturity. For this reason, we have conducted a depth analysis through the existing SaaS maturity model in both business and scientific communities. From Microsoft's model [32], we derived three groups of key benefits, which are related to *efficient and effective services*, *interoperability of services*, and *integrated and secured services*; whereas we found four groups of key benefits within Forrester's model [33], which are *business and IT alignment*, *business and community engagement*, *dynamic and interoperable services*, and *agile delivery*. Kang's model [34] resulted in three groups of key benefits, which are *integrated services*, *effective and efficient services*, and *SOA benefits*.

According to this analysis, six categorizations for the perceived benefits can be developed within SaaS maturity model. They are *Level 0-Potential*, *Level 1-Pro prospective*, *Level 2-Promising*, *Level 3-Achieving*, *Level 4-Stabilizing*, and *Level 5-Optimizing* which can represent different levels of SaaS maturity model that shows what extent benefits of SaaS can be expected by SaaS users.

The *Level 0-Potential* represents limited benefits which can be indicated by utility-based characteristic from pay-per-use or pay-per-period. This maturity level represents an early adoption of SaaS model that resembles application service provider (ASP) model through which traditional client-server application is delivered. Since the application is outsourced to service providers, SaaS users can transfer the risk of managing IT-related function externally. Apart from the early adoption, the *Level 0-Potential* shows a growing concern toward SaaS model and potential benefits which may be expected from SaaS adoption.

The *Level 1-Pro prospective* represents benefits which can be indicated by SLA-driven characteristic that can be shown by standardized services for multiple customers. In this level of maturity, SaaS users are served with services which are standardized within SLAs that are important to measure service performance. Apart from limited benefits that may be expected in this level, the *Level 1-Pro prospective* delineates standardized supports for customized application from service providers during period of contract.

The *Level 2-Promising* represents benefits which can be indicated by resource sharing characteristic that can be shown by efficient management of resources and effective management of business process. SaaS users are allowed to configure SaaS application, but only limited to its interface. The *Level 2-Promising* indicates that an enterprise who adopts packaged application from service providers is ready to enter SaaS market. The enterprise may expect promising benefits from effective and efficient management resources when adopting SaaS application.

The *Level 3-Achieving* represents benefits from which SaaS users may expect reliable services and effective supports toward a traditional SaaS-based application. Infrastructures to support the application, i.e. servers, are shared with multiple customers by service providers. For this reason, reliable services and scalability is of



important concern when delivering the application. The *Level 3-Achieving* indicates benefits from continuous support from services providers to SaaS users.

The *Level 4-Stabilizing* represents benefits by which SaaS users may expect continuous services and configuration options toward SaaS application although they might be limited. Credibility and accountability aspect is of utmost concern due to scalability issues which might arise, e.g. data mixed with other customers when servers are scaled up to many servers. Therefore, in the *Level 4-Stabilizing* benefits of SaaS are represented by credible and accountable services, and collaborative application within which the users may configure the application.

The *Level 5-Optimizing* represents the benefits by which SaaS users may expect the achievement of optimized SaaS benefits from stable but interoperable services. Therefore, in this level of maturity, the benefits are accumulated from the lower level of the maturity model. Services related to SaaS application are continuously improved and sustained to deliver value to SaaS users.

### 3.3 Life Cycle Perspective

Burstein et al. [35] studied about life cycle of SOA and defined sequence of processes which are involved during the delivery. They include *deployment, discovery, composition, selection, mediation, execution, monitoring, compensation, replacement, and auditing*. This life cycle seems appropriate for a SaaS model, although Burstein's works has not yet accommodated continuous improvements. Catalyst Resources [36] suggested a continuous life cycle of SaaS which includes *acquisition* (purchasing and deployment), *installation and setup* (configuration and provisioning), *usage* (support, training, provisioning), and *monitoring and renewal*. The acquisition phase includes evaluation and purchasing of SaaS applications while the installation and setup phase covers installation, deployment and customization of the applications. Within the usage phase, provisioning and training is conducted toward the installed applications which are continuously supported by SaaS providers; whereas monitoring and renewing applications are incorporated within the maintenance and renewal phase.

From the perspective of the life cycle of SaaS, the benefits of SaaS are conceptualized by considering the design principles of SaaS which are appreciated during SaaS life cycle. The life cycle should show both perspectives of providers and users during the delivery of SaaS applications as depicted in Figure 1. According to this figure, the perspective is shown by the parallelogram-shape, while three main phases within SaaS life cycle are outlined within the pointless square-shape. These three main phases are *service discovery, service selection and engagement, and service enactment* which are depicted by the square-shape with bold line. The square-shape represents processes in each phase of the life cycle. The hexagon-shape serves inputs for the processes. There are no significance differences between striped-arrows and bold-arrows unless the later are used to connect phases within SaaS life cycle.

Figure 1 can show that both of these perspectives will be accommodated during the whole life cycle and thus the perspective of SaaS life cycle can be accommodated when identifying benefits of SaaS from the perspective of SaaS providers and users.

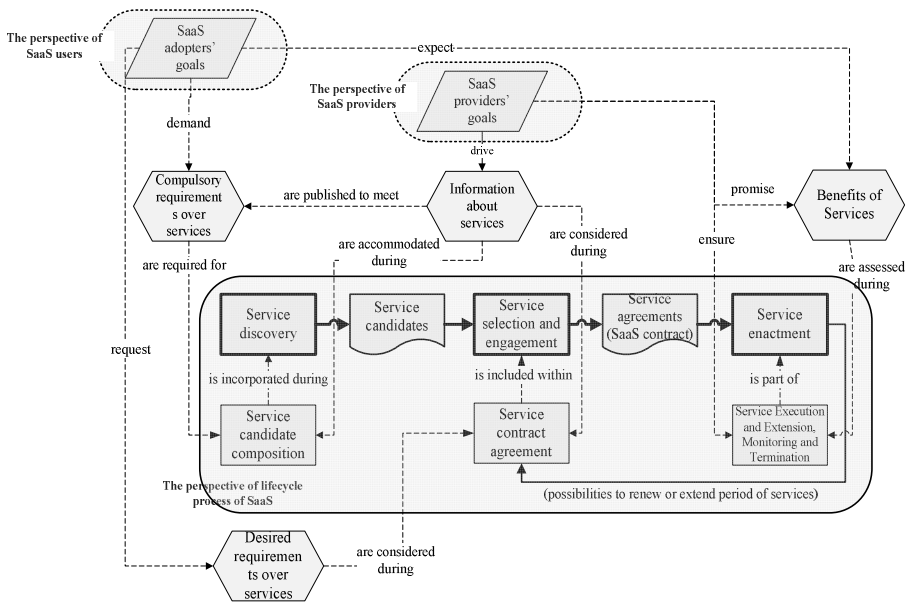


Fig. 1. SaaS life cycle and provider’s and user’s perspective (based on Bursteijn et al.[35])

### 4 Case Study

In order to make use of the conceptual model which prescribes SaaS benefits, we employed a case study to understand SaaS benefits from the perspective of providers, and users. We used Exact Online software as our case study which is financial management software for SMEs. The research data was gathered through semi-structured interview method that are commonly used for qualitative research which aims at understanding respondents’ point of view prior to specific issues which are incorporated within an interview guide which is prepared ahead of time before interviews are conducted.

After identifying data from the interviews, we found some of the promised benefits are fulfilled. Among these benefits, interactive support benefits are of key benefits which are delivered through the adoption of the application. Exact Software Nederland guarantees high security standard to protect client’s data by providing a certified-secured access (e-sure) toward the applications. In their strategy comprehensive support benefits are delivered by providing collaborations with many stakeholders who are involved, including accounting firms, suppliers, banks, with tax-office. Benefits are also be expected by including changes to legislation requiring continuous updates to comply with laws and regulations. Exact Online software can also deliver technology-enabled business support benefits by introducing an efficient and effective financial management to support business process of an enterprise, while the enterprise can focus on its core business process.

Pertaining to the perceived benefits, users of Exact Online software may expect benefits at SaaS maturity *Level 3 (Achieving)* from reliable services which are substantiated by a certified-secured access (*e-sure*) for the software. The case shows that also the benefits from the lower phases are accomplished. Furthermore, Exact also provides limited configuration options for users to change the application interface while scaling data to bigger servers is also made possible. Thus, users may also expect benefits from flexible and scalable services which are substantiated with configurability and scalability attributes that can be found in Exact Online software.

**Table 1.** SaaS benefits from the SaaS provider and user perspectives

Benefits of SaaS	Perspectives	SaaS provider	SaaS user
<b><u>The Promised Benefits</u></b>			
<ul style="list-style-type: none"> <li>• Interactive support</li> <li>• Comprehensive support</li> <li>• Technology-enabled business support</li> </ul>		certified-secured access links to many entities up-to-date with laws and accounting standard	
<b><u>The Perceived Benefits</u></b>			
<ul style="list-style-type: none"> <li>• Subscription-based (Level 0)</li> <li>• Standardized services (Level 1)</li> <li>• Resource sharing (Level 2)</li> <li>• Continuous supports (Level 3)</li> <li>• Flexible and scalable (Level 3)</li> </ul>			various types of subscription module, extensions, add-ons scale up to bigger servers workday supports configurable interfaces

We summarized the promised benefits and the perceived benefits based on our case study in Table 1. According to this table, two important perspectives are accommodated from which benefits of SaaS are addressed by referring to the conceptual model which can help to indentify the benefits. By using the conceptual model, benefits of SaaS can be categorized into the related perspective, and further can be identified with a certain characteristic. A discussion between promised, perceived, and actual benefits was initiated, in which it was found that these were very similar, given the maturity phase. The maturity phase is an important variable for explaining the number of benefits gained in this case study.

During the whole life cycle of SaaS, we also found that the process benefits can be appreciated when delivering Exact Online software as a SaaS application: During service discovery, Exact should provide sufficient information about services to potential customers while consider possibilities of vendor diversification options. During service selection and engagement, Exact should consider collaboration in conceptualizing level of services while also consider possibilities to accommodate business changes. During service enactment, Exact should facilitate accessible and continuous services during period of contract while also consider possibilities to provide reusable and composable services in order to address changes in user requirements and technology.

## 5 Conclusion and Future Work

In this research, different perspectives when realizing SaaS benefits are considered. We conceptualized the benefits of SaaS in each perspective by considering different benefits during different phases of SaaS life cycle and tested the model in a case study. An important finding is that SaaS providers and users might have different perceptions on benefits. Furthermore the benefits are dependent on the maturity and life cycle support. In the case study we found that SaaS providers ought to appreciate the benefits at relatively early phase of the life cycle, including service discovery, and service selection and engagement. This is due to benefits of SaaS introduced by SaaS providers as part of strategic marketing to attract potential customers and further persuade key advantages in adopting a SaaS application despite of immaturity model of SaaS. On the other hand, SaaS users are likely to appreciate the benefits during the later phase of the life cycle, which includes service selection and engagement, and service enactment. SaaS users ought to see practical benefits from adopting SaaS application, and therefore they would likely to appreciate the benefits when they experience, e.g. an improvement toward business process through effective management of enterprise resources, and an efficient technology support through cost saving efforts.

The research is inherently explorative in nature. Different situations might yield different outcomes. The conceptual model should be seen as a support to help practitioners and researchers to analyze and understand the benefits of SaaS application during the whole life cycle of SaaS. In this way better insight can be gained into the promised, perceived, and actual benefits. Therefore, we hope to enhance the possibilities of an SME to experience more benefits when adopting a SaaS application from SaaS providers.

## References

1. Kremic, T., Tukel, O.I., Rom, W.O.: Outsourcing Decision Support: A Survey of Benefits, Risks, and Decision factors. *Supply Chain Management: An International Journal* 11(6), 467–482 (2006)
2. Holcomb, T.R., Hitt, M.A.: Toward a Model of Strategic Outsourcing. *Journal of Operations Management* 25(2), 464–481 (2007)
3. Pettey, C.: Gartner Says 25 Percent of New Business Software Will be Delivered As Software as a Service by 2011. In: *Gartner Symposium/ITXpo 2006* (2006), <http://www.gartner.com/it/page.jsp?id=496886>
4. Dubey, A., Wagle, D.: Delivering Software as a Service. *McKinsey Quarterly* 6, 1–12 (2007)
5. Mertz, S.A., Eschinger, C., Eid, T., Huang, H.H., Pang, C., Pring, B.: Market Trends: Software as a Service. In: *Worldwide, 2008-2013, Gartner* (2009)
6. Gonçalves, V., Ballon, P.: Adding Value to the Network: Mobile Operators' Experiments with Software-as-a-Service and Platform-as-a-Service Models. *Telematics and Informatics* 28(1), 12–21 (2011)

7. Carraro, G., Chong, F.: Software as a Service (SaaS): An Enterprise Perspective. Microsoft Corporation (2006), <http://msdn.microsoft.com/en-us/library/aa905332.aspx>
8. Bleicher, P.: Solutions Delivered, Not Installed. Applied Clinical Trials (2006), <http://appliedclinicaltrialsonline.findpharma.com/appliedclinicaltrials/IT+Articles/Solutions-Delivered-Not-Installed/ArticleStandard/Article/detail/334567?contextCategoryId=554>
9. Software & Information Industry Association.: Software as a Service: Strategic Backgrounder. Software & Information Industry Association (2001), <http://www.siiia.net/estore/ssb-01.pdf>
10. Benlian, A.: A Transaction Cost Theoretical Analysis of Software-as-a-Service (SaaS)-based Sourcing in SMBs and Enterprises. In: Proceedings of the 17th European Conference on Information Systems, Verona, Italy (2009)
11. Benlian, A., Hess, T.: The Risks of Sourcing Software as a Service – An Empirical Analysis of Adopters and Non-adopters. In: Proceedings of the 18th European Conference on Information Systems, ECIS (2010)
12. Kemp Little LLP.: Key Issues in SaaS Contracts (2010), <http://www.kemplittle.com/html/stay-posted/publications/short-lines/key-issues-in-saas-contracts-june-2010.html>
13. Gruman, G.: Include Management Costs When Calculating Software as a Service Benefits (2007), <http://www.cio.com/article/print/111151>
14. Zucco, J.: Benefits of a Software as a Service Model (2006), <http://searchenterprisewan.techtarget.com/feature/Benefits-of-a-software-as-a-service-model>
15. Sun, W., Zhang, K., Chen, S.-K., Zhang, X., Liang, H.: Software as a Service: An Integration Perspective. In: Krämer, B.J., Lin, K.-J., Narasimhan, P. (eds.) ICSOC 2007. LNCS, vol. 4749, pp. 558–569. Springer, Heidelberg (2007)
16. Olsen, R.: Transitioning to Software as a Service: Realigning Software Engineering Practices with the New Business Model. In: International Conference on Service Operations and Logistics, and Informatics, Shanghai, China, June 21–23, pp. 266–271. IEEE (2006)
17. Greschler, D., Mangan, T.: Networking Lessons in Delivering Software as a Service – Part I. International Journal of Network Management 12(5), 317–321 (2002)
18. Choudhary, V.: Comparison of Software Quality Under Perpetual Licensing and Software as a Service. Journal of Management Information Systems 24(2), 141–165 (2007)
19. Aulbach, S., Grust, T., Jacobs, D., Kemper, A., Rittinger, J.: Multi-tenant Databases for Software as a Service: Schema-Mapping Techniques. In: Wang, J. (ed.) Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, June 9–12, pp. 1195–1206. ACM (2008)
20. Xin, M., Levina, N.: Software-as-a-Service Model: Elaborating client-side Adoption Factors. In: Boland, R., Limayem, M., Pentland, B. (eds.) Proceedings of the 29th International Conference on Information Systems, Paris, France, December 14–17. SSRN (2008), <http://ssrn.com/abstract=1319488>
21. Laplante, P.A., Costello, T., Singh, P., Bindiganavile, S., Landon, M.: The Who, What, Why, Where, and When of IT Outsourcing. IT Professional 6(1), 19–23 (2004)

22. Janssen, M., Joha, A.: Challenges for Adopting Cloud-based Software as a Service (SaaS) in the Public Sector. In: European Conference on Information Systems, ECIS 2011, paper 80 (2011), <http://aisel.aisnet.org/ecis2011/80/>
23. Stuckenberg, S., Heinzl, A.: The Impact of the Software-as-a-Service Concept on the Underlying Software and Service Development Processes. In: Proceedings of Pacific Asia Conference on Information Systems, PACIS 2010, pp. 1297–1308 (2010)
24. Nassif, A.B., Capretz, A.M.: Moving from SaaS Applications toward SOA services. In: IEEE 6th World Congress on Services, pp. 187–188 (2010)
25. Watson, R.: Optimizing Cloud Application Architecture with SOA Principles. In: Gartner Webinar (2010), [http://www.gartner.com/it/content/1417900/1417916/september\\_8\\_optimizing\\_cloud\\_application\\_architecture.pdf](http://www.gartner.com/it/content/1417900/1417916/september_8_optimizing_cloud_application_architecture.pdf)
26. Dippenaar, J.F.: Software as a Service (SaaS): Consideration and Implications for SaaS Customers. University of Stellenbosch, South Africa (2008), [https://ir1.sun.ac.za/bitstream/handle/10019.1/15037/dippenaar\\_software\\_2008.pdf?sequence=1](https://ir1.sun.ac.za/bitstream/handle/10019.1/15037/dippenaar_software_2008.pdf?sequence=1)
27. Zucco, J.: Benefits of a Software as a Service Model (2006), <http://searchenterprise.wan.techtarget.com/feature/Benefits-of-a-software-as-a-service-model>
28. SOA Systems, Inc.: Goals and Benefits of Service Oriented Computing. (2011), <http://www.whatissoa.com/pl6.php>
29. Erl, T.: SOA: principles of service design. Pearson Education, New Jersey (2007)
30. Prabowo, A.H.: A Methodology for Assessing the Benefits of Software as a Service: Perspectives and Benefits when Delivering Enterprise Resource Planning as Service within Small and Medium Sized Enterprises. Delft University of Technology, Delft (2012)
31. IT Governance Institute. COBIT 4.1. USA (2007)
32. Carraro, G., Chong, F.: Architecture Strategies for Catching the Long Tail. Microsoft Corporation (2006), <http://msdn.microsoft.com/en-us/library/aa479069.aspx>
33. Ried, S.: Forrester's SaaS Maturity Model: Transforming Vendor Strategy while Managing Customer Expectations (Excerpt Online). Forrester (2008), [http://www.forrester.com/rb/Research/forresters\\_saas\\_maturity\\_model/q/id/46817/t/2](http://www.forrester.com/rb/Research/forresters_saas_maturity_model/q/id/46817/t/2)
34. Kang, S., Myung, J., Yeon, J., Ha, S.-W., Cho, T., Chung, J.-M., Lee, S.-G.: A General Maturity Model and Reference Architecture for SaaS Service. In: Kitagawa, H., Ishikawa, Y., Li, Q., Watanabe, C. (eds.) DASFAA 2010, Part II. LNCS, vol. 5982, pp. 337–346. Springer, Heidelberg (2010)
35. Burstein, M., et al.: A Semantic Web Services Architecture. IEEE Internet Computing 9(5), 72–81 (2005)
36. Catalyst Resources.: Design SaaS to Manage Full Customer Lifecycle User Experience. Business Management (2011), <http://www.bme.eu.com/article/Design-SaaS-to-manage-the-full-customer-lifecycle-user-experience/>

# Collaboration Infrastructure for the Learning Organization

Keith Harrison-Broninski<sup>1</sup> and Janne J. Korhonen<sup>2</sup>

<sup>1</sup> Role Modellers Limited, Somerset, UK  
k**hb**@rolemodellers.com

<sup>2</sup> Aalto University, Espoo, Finland  
j**anne.korhonen**@aalto.fi

**Abstract.** Human Interaction Management (HIM) is a holistic theory of human collaborative work that provides management principles and patterns for business processes focused on knowledge work. The Human Interaction Management System (HIMS) is the associated software technology for process design, execution and management. Goal-Oriented Organization Design (GOOD) is the associated change management methodology. In this paper, we suggest that HIM, the HIMS, and GOOD provide the basis for a collaboration infrastructure that is conducive to the Learning Organization and that exemplifies good socio-technical design.

**Keywords:** Human Interaction Management (HIM), Human Interaction Management System (HIMS), Goal-Oriented Organization Design (GOOD), learning organization, socio-technical design.

## 1 Introduction

Most knowledge workers find it hard to visualize and manage the structure of their work. They also do not know to what extent they are efficient or effective. In fact, research shows that they are neither. Multi-year time and motion studies on knowledge workers in the US [15] show that they spend an average of 28% of their working day organizing their interactions with one another rather than doing useful work. The cost of this wasted time to employers, and hence to the US economy, is estimated to be 650 billion dollars per annum, which equates to a cost to the worldwide economy of something like 2 trillion dollars per annum. So knowledge workers are not doing things right. Are they even doing the right things? As an example of how ineffective knowledge workers are, further research shows that although 82% of all organizations are undertaking some form of change initiative at any one time [2], 70% of change initiatives fail [4].

Traditional management techniques for describing work arose in manufacturing. In the Scientific Management [16], outputs are specified in advance, a sequence of tasks to deliver them is defined, and the tasks are then scheduled. We view that this mechanistic approach is increasingly outdated in the face of the mounting complexity of the business environment. While it is widely

acknowledged that knowledge economy calls for “learning organizations” that continuously adjust to changing requirements and priorities, the management tools and techniques are still predominantly rigid and geared to relatively stable and predictable business environments. Technological and methodological support for collaborative knowledge work and organizational learning is rather incipient. Prevalent knowledge management and groupware tools have generally made communication faster, but they have not been able to make human collaboration more effective. In this paper, we will look into a proposed alternative approach that provides collaboration with a process context: Human Interaction Management (HIM) [7]. HIM is a holistic theory of collaborative human work that draws from psychology, biology, social systems theory and learning theory.

HIM provides a set of principles and patterns for designing, executing and managing business processes focused on knowledge work. Whereas the mainstream Business Process Management (BPM) techniques and tools deal with “mechanistic” business processes, in which human involvement is limited to key data entry and decision points, HIM extends work support to “human-driven” processes focused on human creativity and collaboration. HIM provides process-based support for innovative, adaptive, collaborative human work and allows it to be integrated in a structured way with routine work processes that are often largely automated via BPM systems or other technologies.

A Human Interaction Management System (HIMS) is a process modeling and execution tool based on HIM theory: “process modeling and enactment system that provides native support for the six Role Activity Theory object types (Role, Entity, Activity, User, State and Interaction), uses a state-based approach to Activity enablement and validation, permits Interactions to be composed of multiple asynchronous channels, and supports management of process change by allowing any process component to be created and configured as a natural part of process execution”. [7, p. 266]. The reference implementation of a Human Interaction Management System (HIMS) is the software HumanEdj, which we will cover in more detail in Section 3.

HIM has an associated methodology called Goal-Oriented Organization Design, or GOOD [8]. GOOD starts with a Design Stage in which organizational domain is defined, vision and mission are designed and refined, stakeholders are identified, and benefits profiled. The next Stage, Delivery, includes requirements analysis, stakeholder management, risk/dependency management, and operational transition as a unified whole. The final Stage, Optimization, ensures that communications are effective and benefits managed. In practice, all three Stages may run in parallel for much of the time.

We view that together HIM and GOOD provide a collaboration infrastructure that formally supports management of knowledge work in learning organizations.

The paper is structured as follows. In Section 2, we discuss the theoretical underpinnings of HIM, in terms of the inner structural patterns of collaborative human work. In Section 3, we discuss the practical implementation of HIM as a HIMS. Based on the findings in these sections, we suggest in Section 4 that HIM provides a collaboration infrastructure that, with the associated GOOD



methodology, is conducive to the Learning Organization. In Section 5, we show that HIM and GOOD together exemplify good socio-technical design. In Section 6, we recount a case study of applying HIM in an innovation organization. Finally, Section 7 provides a brief discussion and concluding remarks.

## 2 Human Interaction Management Theory

Drawing from and extending Role Activity Theory [10], Human Interaction Management theory provides a modelling framework for describing collaborative human working behaviour in process terms, and identifies patterns that underlie any form of human activity (whether collaborative or not) and demonstrate how learning is the core of all collaborative work [8]. Harrison-Broninski [7] starts the analysis of human-driven processes from the inner structural patterns rather than from their external manifestation in terms of particular communications. He suggests that any human work consists of five stages [7]:

- *Research*. This stage is about mapping out the terrain of the work; gaining information from external sources, e.g. communities of practice, textbooks, Web search, and turning it into personal knowledge.
- *Evaluate*. Here one steps back to consider and internalize the acquired knowledge.
- *Analyze*. An approach to the problem is decided upon, at least initially.
- *Constrain*. The work is divided into separate chunks and organized. This stage is about laying down the constraints that govern these chunks of work.
- *Task*. As the chunks of work have been handed out to appropriate people, all those concerned can get on with the tasks at hand.

The first stage of the REACT pattern, Research, is further broken down into a sub-pattern AIM, which describes the activities of information discovery:

- *Access* discovery services. This stage is about mapping out the terrain of the work; gaining information from external sources, e.g. communities of practice, textbooks, Web search, and turning it into personal knowledge.
- *Identify* resources required. At this sub-stage, resources of likely interest and usefulness are identified and chosen.
- *Memorize* information obtained from particular resources. This sub-stage is about internalizing the ideas in question.

According to Harrison-Broninski [7], the REACT and AIM patterns describe all human working behaviour. The patterns capture the way that people react to the work they take on: e.g. respond to an assignment, fulfil a responsibility, achieve a goal. REACT and AIM help simplify complex situations since the patterns can be repeated, overlapped, and nested in order to reduce any work assignment to the same fundamental stages.

### 3 The Human Interaction Management System

Implementation of HIM in an enterprise environment (i.e., design, execution and management of business processes according to HIM principles) is facilitated by software support from a Human Interaction Management System (HIMS). The aim of a HIMS is to facilitate all stages of human work without forcing people to follow a set of predetermined steps. A HIMS helps people to see the bigger picture of a process and understand their responsibilities within it. This calls for suggestive rather than prescriptive process description and support: a HIMS provides support and enforces basic control on behalf of the organization, providing an indication to people of what they are expected to do then letting them learn collaboratively how best to meet their assigned goals [7].

A key aspect of this collaborative learning derives from autopoietic theory [18], which asserts that communication is founded not on transmission of information but rather on transmission of intent. Research in biology shows that the purpose of animal communication is largely about synchronizing the behaviour of parties. This understanding has been adopted in business via the classic “Conversation for Action” pattern, in which communication between people and organizations is structured in terms of a small set of request/response pairs – request/promise, offer/acceptance, report/acknowledgement [7]. HIM generalizes this principle by allowing a much broader and less restrictive set of structured communications.

HumanEdj software, for instance, provides full support for speech acts theory [7], according to which a communication act is not only composed of content but also, and at least as importantly, of an intention. A speech act lets the sender of a message specify two things about the message [13]:

- The *Intended Manner*, or the illocutionary force, that describes the tone of voice one is adopting: for instance “Advise” rather than “Require” in order to make it clear that the message is a suggestion rather than an order.
- The *Intended Effect*, or the performative, that describes the sort of thing one wants to happen as a result of the message: for instance “Ask\_All” rather than “Ask\_One” in order to make it clear that one wants the recipient(s) to canvas their entire team about something, rather than just key team members.

Many business people have found the traditional use of speech acts in the “Conversation for Action” too rigid for practical use [5]. Hence, HumanEdj permits business people not only to share data and documents, but also to make a wide range of assertions about the status of Deliverables and Stages. More generally, a HIMS suggests actions rather than prescribes them, allows not only for communication but also for action, does not assume that all communication is direct and does not prevent tangential discussion, i.e. unexpected interactions that go beyond the conversation originally expected [7]. This permits processes to evolve via a collaborative learning process. Rather than being based on a specific aspect of human collaboration such as speech acts, a HIMS is based on the five fundamental features of human-driven processes identified by Harrison-Broninski [7]:

1. *Connection visibility.* Collaborative technology must provide a strong representation of process participants, their roles and their private information resources. To work with people, one needs to know who they are, what they can do, and what their responsibilities are.
2. *Structured messaging.* If people are to manage their interactions with others better, their communications must be structured, goal-directed and under process control.
3. *Support for mental work.* Human-driven process support must recognize the value of the human information processing: the time and mental effort invested in researching, comparing, considering, deciding, and generally turning information into knowledge and ideas.
4. *Supportive rather than prescriptive activity management.* People may not sequence their activities in the manner of a software program, but there is always structure to human work, which must be understood and institutionalized so that it can be managed and improved.
5. *Processes change processes.* Process definition is an intrinsic part of the process itself; it happens continually throughout the life of the process.

The HIMS imposes structure by modelling work formally as a process. By bringing collaboration tools into a unified process context, it promises to make work genuinely more effective [9].

Situated learning [12] suggests that all learning is contextual, embedded in a social and physical environment. Personal knowledge and problem solving are closely tied up with interrelations with others and the artefacts used. This leads to communities of practice [20], in which a group of people, bound together by informal relations, naturally develops common language, understanding of their work context and meaning attached to their tools. The HIMS can be seen as facilitating socialization into a community of practice by providing “scaffolds” [1] in the “zone of proximal development” [19]. As Hutchins puts it, the longer-term transmission of knowledge is “crystallized and saved in the physical and conceptual tools of the trade and in the social organization of work” [11].

The reference implementation of a HIMS, which is arguably most true to the HIM theory and thus our software of choice for the case study, is HumanEdj [6], which is free for small-scale use. HumanEdj has a distributed peer-to-peer architecture, more akin to a Multi-Agent System than to a workflow engine. Participants in a process, which in HumanEdj is called a “Plan” may belong to different organizations and use different HumanEdj instances. HumanEdj automatically synchronizes the Plan state for all participants via a messaging technology such as email. It is also possible to participate in a Plan using a standard email client.

HumanEdj structures activities, messages, documents and data as well as maintains information on who does what, when, where and why. Fine-grained control over who sees what in a Plan is accomplished by grouping all the above items into “Stages”, each of which represents a related set of goals and effectively defines a virtual sub-team within the Plan. Plans may generate sub-Plans, for instance in order to carry out the details of a public process inside distinct private processes [8].

Plan templates are used to generate Plans for projects, initiatives, ventures, etc. – i.e. executable business processes that may cross-organizational boundaries. Each Plan is configured appropriately for the requirements of the situation. The participants themselves adjust the configuration throughout its life, as they collaborate to evolve the definition of the Plan instance in response to external circumstances and internal progress.

A Plan instance acts not only as a mechanism for learning but, once complete, as a source of learning materials. Plan instances from a repository show how other people dealt with problems of a certain type, and new Plan templates may be created from successful Plan instances (or parts thereof).

With regard to assessment of learning results, Plan instances are self-monitoring – they include automatic feedback mechanisms both within the Plan and across Plans to higher management levels. Taking part in a Plan instance in itself both measures and provides evidence of achievement. Plans may also use external services to provide:

- Learning materials customized for the Plan instance
- Standardized evaluation of learning progress
- Trusted competency assessment
- User profiles

Information within a Plan instance automatically has semantic mark-up, as do all communications between participants. This mark-up can be sent to external services to help streamline the results.

## 4 HIM and the Learning Organization

Collaboration is fundamental to what is called a Learning Organization [14]. A Learning Organization facilitates the development of skills and experience by its members and continuously transforms itself via on-going negotiation between its members. HIM and the associated GOOD methodology allow the organization to structure work around learning, resulting not only in a more fulfilling workplace for the individual but also in improved performance for the organization and its partners. HIM and GOOD provide a way to define, implement, monitor and adjust organizational goals and strategies – integrating different levels of management both within and across organizations into a dynamic process network driven by learning.

In their meta-analysis of themes in the learning organization literature, Thomas and Allen [17] synthesized five broad categories:

1. *Learning*. The nature of learning at the individual level, its effect and application at the team level, and amplification at the organizational level. This is empowered by the space and recognition that HIM mandates for mental work in collaborative human activity (for example, HumanEdj Plan templates typically have deliverables that represent the concrete outputs of thought and discussion over a period of time), and the emphasis on such

work in the REACT pattern. Of the five Stages of REACT, the first three are entirely devoted to learning, and the first Stage (Research) is given a special emphasis by being separated into the three sub-Stages of the AIM pattern.

2. *Structure.* The basis and composition that enable the organizational learning processes and systems. This is provided by the re-use of Plans as Plan templates, in whole or part, for particular Plans that embody solutions to special cases or new situations.
3. *Shared vision.* The binding component and catalyst of organizational change. This is assured by the use of Stages to represent sets of related goals and Roles to represent corresponding responsibilities. By making goals and responsibilities concrete, they become visible to participants in a Plan, who can then discuss and negotiate them to ensure shared understanding.
4. *Knowledge management.* The capture, structuring and re-conceptualization of implicit and explicit knowledge. This is given a practical basis in the re-use of Plan templates described above, since they make the solutions derived in practice explicit, as instructions for future operations.
5. *Strategy.* By which the organization identifies potential and capitalizes on the opportunities. This is dealt with via the GOOD method, which provides a standard, universal set of Stages, Roles, Activities and Deliverables to manage the complexity of organizational change.

## 5 HIM and Socio-technical Design

Cherns [3] provides a basic framework for understanding and designing socio-technical systems in consideration of human and social aspects. Building upon the notion of a participative process, he defines nine key principles of socio-technical design. It appears that HIM and GOOD are congruent with these principles and thus representative of good socio-technical design:

1. *Compatibility.* The process of design must be compatible with its objectives. HIMs objective is to support irregular collaborations and give them an appropriate process context. Likewise, the process of design is a collaborative effort by the people taking part in the managed processes, and is focused on defining and sharing sets of related goals.
2. *Minimal Critical Specification.* No more should be specified than is absolutely essential, yet what is essential will be identified. Process description and support in HIM are suggestive rather than prescriptive: a HIMS provides support and enforces control, advising people on what to do and letting them carry out their tasks as they see fit.
3. *The Socio-technical Criterion.* Variances, if they cannot be eliminated, must be controlled as near to their point of origin as possible. In HIM, each participant in the system is responsible for executing their private process and accountable for others as specified in the public process.
4. *The Multifunctional Principle.* As opposed to the traditional form of organization, in which people perform highly specialized and fractionated tasks,

Cherns calls for multifunctional and equifinal mechanisms that can provide a range of responses by using different combinations of elements. HIM allows the network “wiring” between the participants to change and enables fluid behaviour of processes in the declarative bounds of specified channels. The HIMS suggests actions rather than prescribes them and supports unexpected interactions.

5. *Boundary Location.* Departmental boundaries interfere with desirable sharing of knowledge and experience. The role of the manager should be concentrated on the boundary activities: ensuring adequate resources, coordinating with other departments, etc. HIM terms this executive control: determining the Roles, interactions and deliverables of a process.
6. *Information Flow.* Information systems should supply people with exactly the right amount of information to enable them to control the variances that take place in their sphere of responsibility and competence. HIM provides each Role with access to its own data, sharing it with other Roles only through message exchange on an as-needed basis and always within a purposeful context (a goal-directed Stage).
7. *Support Congruence.* The systems of social support should be congruent with the behaviours, which the organization structure is designed to elicit. If the intention is to improve collaboration and increase organizational effectiveness via learning, HIMS provides the right kind of means and constraints to achieve this since the five principles of HIM are based on deep understanding of human interactions and place high value on learning as an aspect of collaborative behaviour.
8. *Design and Human Values.* The outcome of organizational design should be a high quality of work. HIM does not limit human involvement to key data entry and decision points, but also makes goals and responsibilities explicit and supports the corresponding human interactions, thereby encouraging and leveraging purposeful, skilful decision-making and judgment.
9. *Incompletion.* Design is a reiterative process: “As soon as design is implemented, its consequences indicate the need for redesign.” HIM enables continual change (renegotiation) of processes on the fly.

## 6 Case Study – An Innovation Organization

To illustrate how HIM supports collaboration and organizational learning, we will consider an innovative company whose products are improvement programmes that it delivers to public sector organizations. The management structure is flat and staff members are encouraged to propose, seek internal funding for, and implement new improvement programmes on a regular basis. While the culture has resulted in innovations beneficial to their customers, and consequently in growth, the company struggled to make its operations profitable. It was not possible to optimize or even obtain the cost of sales, given the complex way in which improvement programmes were created, sold, and delivered. It became necessary to standardize and monitor customer-facing operations.

The company expected to continue its previous success with standardizing back-office administrative processes using traditional workflow techniques. However, standardization of customer-facing operational processes met with resistance from staff, who were accustomed to using their skills, experience and judgement to adapt their working approach to each customer engagement. Hence, there remained wide variance across the organization in the way that core customer-facing and internal processes were carried out.

The solution required a means of process standardization that provided indicative rather than prescriptive processes (i.e. processes that could be adapted flexibly during execution), and that supported the harvesting of innovative ideas into new products (i.e. improvement programmes). The company used HIM to develop Plan templates for core operational processes including:

- *Sales Funnel*. Developing a sales lead into a new customer engagement.
- *Product Delivery*. Implementing an improvement programme for a customer.
- *Non-Standard Product Development*. Developing a custom improvement programme for a customer.
- *Standard Product Development*. Turning a custom improvement programme into a standard off-the-shelf product offered to all customers.

Shown below in Fig. 11 is a HumanEdj “Grid view” of the Plan template for the Sales Funnel process. Across the top are the Roles in the process, which in an actual Plan would be assigned to named people. Down the side are the Stages in the Plan template – the numbering is only suggestive, since the Stages may be carried out in any order, and they often run concurrently. Here we see Shared vision [17] made concrete via use of Stages to represent sets of related goals.

During the lifetime of a Plan, the Stages will be assigned statuses by the Plan owner, such as “Started”, “Completed”, “Cancelled”, and so on. Different Roles belong to different Stages. Any documents, data or messages created in a Stage are visible to all the Roles in that Stage and only to those Roles. Here we see the emphasis on mental work that is critical to Learning [17], via deliverables identified and recognized as a natural part of Plan execution.

Two Activities in particular are to be noted:

1. “Initiate Non-Standard Product Development” in Stage “Develop Opportunity”, which involves the creation of a new sub-Plan for developing a custom improvement programme, if required. The sub-Plan will be based on a standard Plan template, adapted as required. If the standard Plan template is adapted, the new version may itself become a standard Plan template for use by others. The creation of the sub-Plan not only draws on organizational knowledge about custom improvement programme creation, but may well contribute to it by addition of a new special case. Here we see how the creation of a particular sales proposal contributes to evolving organizational Structure [17], since the way in which it was done is automatically made part of enterprise Knowledge management [17].

Somerset GP Service Q4 2011 :: 31-Oct-2011 14:05:29.390

Stage	Nominated Sales Lead	Lead Owner	Lead Creator	Client	Solutions Team	Product Specialist
<b>1. Generate Lead</b>	View Lead in CRM (-17 days, due 04-Nov-2011)	View Lead in CRM (-14 days, due 09-Nov-2011)	Maintain Lead in CRM	Not in Stage		Not in Stage
<b>2. Qualify Lead</b>	Qualify Lead	View Qualified Lead (-8 days, due 17-Nov-2011)	View Qualified Lead	Not in Stage	View Qualified Lead	Not in Stage
<b>3. Create Opportunity</b>	Assess Client Arrange Follow-Up Meeting Record Change to Opportunity on CRM	View Pre-Meeting Document	Not in Stage	Not in Stage	Not in Stage	Not in Stage
<b>4. Develop Opportunity</b>	Proposal	Approve Proposal Submission	Not in Stage	Not in Stage	Not in Stage	Review Non-Standard Product Offering
	Formal tender					Initiate Non-Standard Product Development
	Submit Proposal					
<b>5. Negotiate Proposal</b>	Send Proposal to Client			Review Proposal		Not in Stage
<b>6. Await Decision</b>	Prepare for Delivery		Not in Stage	Not in Stage	Not in Stage	Not in Stage
	Initiate Delivery Plan					
<b>7. Close Opportunity</b>	Create Contract Close Opportunity in CRM Initiate Resources Allocation For Delivery	View Opportunity Status	View Opportunity Status	Not in Stage	Not in Stage	Not in Stage

**Fig. 1.** Excerpt from HumanEdj Grid View in tabular format of Plan template for sales of improvement programmes

- “Initiate Delivery Plan” in Stage “Await Decision”, which involves the creation of a new sub-Plan for delivering the improvement programme. The Plan template used for this is created as part of the proposal and adapted for each customer engagement. As above, creation of a sub-Plan for a particular Delivery may well result in an adapted Plan template that can be re-used for future Deliveries of the same type. This creation of one Plan from another is typical of the GOOD method, which can be used at any level in an organization to align operations with Strategy [17](#).

Statistics from the Delivery sub-Plan are used together with statistics from the Sales Funnel Plan itself (shown for an example template in Fig. [2](#)) and any sub-Plan for Non-Standard Product Development to generate accurate total cost for provision of the improvement programme to the customer, and hence to create a price that ensures the engagement returns a profit (or a deliberate loss).

By explicitly associating the different aspects of customer engagement with one another, the organization is making its customer-facing operations and their internal relations *visible*. This means not only that senior management can learn to manage the processes as a unified whole, but also that new staff can learn what the organization actually does and how they fit into it. These means of learning are fundamental enablers as the organization grows, since geographical expansion means that teams are increasingly virtual and operational staff includes more and more sub-contractors rather than employees.



	Work To Do	Effort Days	Effort Cost - Total	Effort Cost - Remaining Work	Earliest Activity Start Date	Latest Activity Deadline	Latest Activity Expected Finish Date	Minimum Activity Expected Margin Days				
Plan	Somerset GP Service Q4 2011 :: 31-Oct-2011	TRUE	11	4,230	4,230	01-Nov-11	17-Nov-11	28-Nov-11	-17			
Description	Instances of this Plan template are created via an Intranet form that anyone can use. The form enters client details into CRM, then starts the Plan pre-populated with: 1. A link to the client page in CRM; 2. A Solutions Area Director assigned to the Lead Owner Role - this Role may be re-assigned during the Plan if necessary.											
ROLE	DAY RATE	OVERHEAD PERCENTAGE	DAY RATE	DESCRIPTION								
Nominated Sales Lead	0	0	402									
Lead Owner	0	0	402	<i>All unqualified leads should be assigned to relevant Solutions team Area Director as the lead owner. The Lead Owner Role may be re-assigned during life of the Plan if necessary.</i>								
Lead Creator	200	0	230	<i>New sales leads that have been generated that are not yet qualified as being a genuine lead are referred to as unqualified leads. These are expressions of interest for our products/services from a variety of sources e.g. simple conversation, email, enquiry in response to marketing/website etc. These leads can be generated and logged by all individuals in the business in this early stage. In some instances the solutions team may ask these individuals to maintain that early relationship and link, until it is appropriate for the solutions team to get involved from a sales perspective. We must be careful that they are leads and not support queries that we log.</i>								
Client	50	0	57									
Solutions Team	0	0	402									
Regional Sales Support	0	0	402	<i>Supports Nominated Sales Lead in preparing the proposal</i>								
Central Sales Support	0	0	402									
Product Specialist	0	0	402									
Area Coordinator	500	0	575									
Business Development Manager	500	0	575									
Defaults	350	<i>15 Defaults are used where not set specifically for a Role</i>										
Stage	Role	Activity	Deliverables	Resources	Work To Do	Effort Days	Effort Cost - Total	Effort Cost - Remaining Work	Start Date	Deadline	Expected Finish Date	Expected Margin Days
1. Generate Lead	Nominated Sales Lead				TRUE	0.4	126.4	126.4	01-Nov-11	09-Nov-11	28-Nov-11	-17

Fig. 2. Excerpt from HumanEdj Summary View in tabular format of Plan template for sales of improvement programmes

Further opportunities include passing on the learning benefits of HIM to client organizations in the form of Plan templates that support their resulting change management initiatives; and use of GOOD to develop the growth strategy. The company has effectively started the latter already, by creating Bottom-Up Plan templates for core operations. The next step is to build a Process Architecture to represent their domain of interest, define vision and mission at multiple levels via a Business Motivation Model, develop understanding of their stakeholders, and create Benefits Profiles for the changes that they plan.

## 7 Discussion and Conclusions

In this paper, we have introduced Human Interaction Management (HIM) [7] and the associated Goal-Oriented Organizational Design (GOOD) method [8]. We have discussed how HIM and GOOD together provide a collaboration infrastructure that formally supports knowledge work, is conducive to the Learning Organization and exemplifies good socio-technical design. In Human Interaction Management, learning is viewed as the basis of all human working activity, recognizing that learning is not only part of all work, but that much work is learning-centred. Whereas traditional scientific management approaches and corresponding workflow tools are inadequate for structuring work around learning, HIM integrates learning into work naturally. GOOD is a methodology for

the creation of organizations that are empowered by learning – organizations in which learning is a driving force for all business processes.

To validate the claims, we have presented a case study of a company that formalized its approach to collaboration and organizational learning with the HIM approach. Although the validation is not conclusive, the case study provided encouraging results: improved visibility into collaborative knowledge work, successful support for indicative rather than prescriptive processes, and natural integration into operational work of all the five themes of Learning Organization [17]. Further research is required to corroborate the tentative findings.

## References

1. Balaban, N.: To become a teacher (1995)
2. Benedict, A.: 2007 change management. Survey Report (2007)
3. Cherno, A.: The principles of sociotechnical design. *Human Relations* 29(8), 783–792 (1976)
4. Dirkin, H.L., Keenan, P., Jackson, A., Kotter, J.P., Beer, M., Nohria, N., Duck, J.D.: Lead change—successfully, 3rd edn. HBR Article Collection (2009)
5. Flores, F., Graves, M., Hartfield, B., Winograd, T.: Computer systems and the design of organizational interaction. *ACM Transactions on Office Information Systems* 6(2), 153–172 (1988)
6. Harrison-Broninski, K.: Introduction to humanedj, <http://rolemodellers.com/resources/HumanEdj.pdf>
7. Harrison-Broninski, K.: Human Interactions: The Heart and Soul of Business Process Management. Meghan-Kiffer Press (2005)
8. Harrison-Broninski, K.: Goal-oriented organization design. *Agile Product & Project Management* (June 2009)
9. Harrison-Broninski, K.: Human interaction management and learning. *BPTrends* (November 2009)
10. Holt, A.W., Ramsey, H.R., Grimes, J.D.: Coordination system technology as the basis for a programming environment. *Electrical Communication* 57(4), 308–314 (1983)
11. Hutchins, E.: *Cognition in the Wild*. MIT Press (1995)
12. Lave, J., Wenger, E.: *Situated Learning: Legitimate Peripheral Participation*. University of Cambridge Press (1991)
13. Role Modellers: Humanedj 3.0.22 Tutorial (March 2008)
14. Senge, P.M.: *The Fifth Discipline*. Doubleday (1990)
15. Spira, J.B., Goldes, D.M.: Information overload: We have met the enemy and he is us. *Tech. rep.*, Basex (2007)
16. Taylor, F.W.: *The Principles of Scientific Management*. Harper and Brothers (1911)
17. Thomas, K., Allen, S.: The learning organisation: a meta-analysis of themes in literature. *The Learning Organization* 13(2), 123–139 (2006)
18. Varela, F.J.: *Principles of Biological Autonomy*. Elsevier (1979)
19. Vygotsky, L.S.: *Mind in Society*. Harvard University Press (1978)
20. Wenger, E.: *Communities of Practice: Learning, Meaning and Identity*. Cambridge University Press (1998)

# IdeaWall: Mixed Mode Distributed Collaboration in Enterprise Environments

Marek Kowalkiewicz

SAP Research  
marek.kowalkiewicz@sap.com

**Abstract.** In early stages of design and modeling, computers and computer applications are often considered an obstacle, rather than a facilitator of the process. Most notably, brainstorming, process modeling with business experts, or development planning, are often performed by a team in front of a whiteboard. While "whiteboarding" is recognized as an effective tool, low-tech solutions that allow remote participants to contribute are still not generally available.

This is a striking observation, considering that vast majority of teams in large organizations are distributed teams. And this has also been one of the key triggers behind the project described in this article, where a team of corporate researchers decided to identify state of the art technologies that could facilitate the scenario mentioned above.

This paper is an account of a research project in the area of enterprise collaboration, with a strong focus on the aspects of human computer interaction in mixed mode environments, especially in areas of collaboration where computers still play a secondary role. It is describing a currently running corporate research project.

**Keywords:** mixed mode collaboration, multimodal collaboration, enterprise applications, mobile applications.

## 1 Introduction

The topic of collaboration in enterprises has been drawing attention of researchers for many years [16]. And while various tools have been made available to employees, the recent rise of interest in "traditional" brainstorming tools, such as whiteboards, post it notes, and other physical artifacts [7], highlights the lack of hardware and software support of distributed collaboration in this particular type of business activities [8].

When looking at the broader topic of distributed collaboration, there are collaboration tools for virtual teams, such as Microsoft NetMeeting or PGI Connect, that can be used in conference calls (synchronous distributed collaboration tools). Support for collaboration extends to asynchronous collaboration as well,

---

<sup>1</sup> See for instance <http://gogamestorm.com> for a comprehensive list of corporate "games" that can be played in front of a whiteboard.

with tools such as JIRA (supporting developer teams), SAP StreamWork (supporting decision makers) or Gravity (used by process modelers). However, early stages of discussion in creative environments often benefit from a pen and packing paper approach. As often witnessed, many information workers prefer to work with traditional brainstorming tools (e.g. Post-it notes, pens and whiteboards) during these early stages. However, if said information workers need to collaborate with remote team members, there is no IT support enabling them to share information and collaborate easily. This means that virtual participants in the meetings (joining using phone or video conferencing tools) are practically blind once this activity starts. This observation has led us to focus our research on IdeaWall technology that allows workers to share and manipulate pen and paper information on a traditional whiteboard with colleagues at one or more locations. With the current version of IdeaWall, undergoing tests at a number of customer sites, each person in the collaboration can be active (i.e. any participant can modify the whiteboard contents) and changes are replicated simultaneously at all locations, creating a so called virtual whiteboard.

When initiating the IdeaWall project, the main question in our technology research was: "How can we let workers use traditional brainstorming tools (e.g. Post-it notes, drawings on a whiteboard) and still let them collaborate and share information with virtual participants in their teams at other locations?"

We also decided to focus on a solution that would not require any custom built or expensive hardware. The requirement was to use hardware that would already be available in a typical meeting room in an organization. This important assumption means that our solution could not use any of the sophisticated tools that are available on the market (such as Microsoft Kinect or Surface), at the same time allowing hardware such as mobile phones, laptop computers, or multimedia projectors. We focused on incorporating state-of-the-art software solutions to implement our vision. That fact differentiates our work from other approaches mixing physical and virtual media [4].

## 2 IdeaWall

IdeaWall allows workers to share and manipulate pen and paper information on a whiteboard with colleagues at other locations. Using a set of IdeaWall Rooms, physical media (e.g. post-it notes, pen markings on a whiteboard) can be placed on a whiteboard in one location and simultaneously visually augmented onto whiteboards at other locations (supported by standard office IT equipment, detailed further in the article). In other words, physical media in one setting can be replicated in one or more other locations and vice versa.

Each IdeaWall Room uses a camera to capture the position and movement of post-it notes, writing, relationships and other physical media placed on the whiteboards. Each capture replaces a physical object with an electronic image equivalent. Objects on the board (either physical or virtual) can be manipulated (e.g. deleted, moved, edited) at any location. Information placed on the walls

can also be stored electronically as long-term documentation. The contents of post-it notes can be captured by a still, high-resolution digital camera.

During our research, we have found out that mobile phones available currently on the market offer computing power and camera resolutions that are sufficient in most typical scenarios. This observation has been crucial in our research, as it was one of the main decision points to focus on a mobile application solution, where all of the client-side functionality of IdeaWall would be available on a mobile phone. The only additional hardware requirement is a video projector, which can be connected to any modern mobile phone. Figure 1 shows a scenario, where two participants of a brainstorm work with a virtual whiteboard while only using a mobile phone and a projector.



**Fig. 1.** Two participants of an IdeaWall brainstorm

The IdeaWall solution allows for a practically unlimited number of remote participants work in front of virtual whiteboard such as in Figure 1. In practice, however, we have not yet worked with customers that would use IdeaWall in a setup with more than three physical locations.

As every virtual whiteboard is stored in our data store as a graph model, together with images of Post-it notes etc., it is also possible to create a standalone web application accessing the models. We have built a web client of IdeaWall, allowing users who - for various reasons - cannot work in front of a physical whiteboard to participate in IdeaWall collaboration as well. To demonstrate the flexibility of such approach, we have embedded IdeaWall whiteboards in SAP

StreamWork - one of the more popular enterprise collaboration tools used by decision makers. The decision to use SAP StreamWork was also driven by the fact that Gravity, or collaborative process modeling tool, is available in SAP StreamWork, and a number of users of Gravity have signaled to us that support of collaboration in front of a whiteboard would be desirable. Figure 2 shows a mobile phone and a portable computer displaying the same whiteboard as shown in Figure 1, this time available to those who prefer computers to traditional brainstorming tools



Fig. 2. IdeaWall available on mobile and portable devices

### 3 Enterprise Applications

In our research, a number of potentially interesting application scenarios has emerged. This section lists several of the more prominent scenarios. The team working on IdeaWall is currently focusing specifically on support of the listed scenarios.

#### 3.1 Early Stages of Business Process Modeling

Business process modelling is an important activity in enterprise management. In the early stages of process modeling, designers use a whiteboard and a set of post-it notes (potentially grouped, potentially linked, but not in a formal way) to define the initial process design. In these early stages, it is important to align views and extract process knowledge from participants. Unfortunately, information workers are not able to use these tools collaboratively author information with colleagues at other locations.

IdeaWall, however, provides a new opportunity in the support of collaborative modeling allowing every participant involved in the collaboration to be active. In other words, any participant in a fully-equipped IdeaWall room can modify the whiteboards design and this gets replicated in all locations. This solves the collaboration problems usually associated with dispersed process design activities in enterprise scenarios.

The later process design phase often occurs in front of a computer, in a BPMN tool such as Gravity. IdeaWall can automatically transfer the results of the collaboration into any BPMN 2.0 compliant system. A business process expert can then take this information and turn it into a proper BPMN model (i.e. post-it notes can be transformed into activities etc.).

### 3.2 Requirements Engineering

Requirements analysis is critical to the success of any software development project. Stakeholders and developers need to be involved in the early stages. However, this becomes complicated when actors are located in different offices, cities, or even countries. IdeaWall, however, can make the requirements elicitation process easier.

Users in different locations can use traditional tools (e.g. whiteboard and markers) to collaboratively author a high-level view of the requirements process. Developers and stakeholders in different locations collaboratively author use case diagrams in UML notation. The results of the collaboration are also captured and imported to a UML tool in real-time.

IdeaWall makes the requirements elicitation process easier, as co-located developers and stakeholders are able to collaboratively model using traditional, customary tools. In addition, temporary whiteboard diagrams are simultaneously stored as long-term documentation which can be kept for later use.

### 3.3 Supply Chain Modeling

While collaboration inside one organisation is already challenging when distributed participants are considered, collaboration among a number of organizations, is even more challenging. A typical situation in a supply chain is one where a potentially large number of stakeholders, through a value chain, contribute towards an end product. In order to maintain flexibility, it is crucial that participants can form ad-hoc business networks and respond to changing market conditions. IdeaWall can help in supply chain modeling, where heterogeneity of information systems of network participants often prevents them from collaborating efficiently. With IdeaWall, the participants can freely model a supply chain using the standard collaboration methods, and after the process has concluded, each participant can import the model into their respective supply chain management systems for further detailing and processing. Especially, interoperability issues such as the business documents, which often need to be discussed outside of the flow diagrams, can be discussed using holodeck.

## 4 Solution Architecture

As already mentioned in Section 2, IdeaWall in general can be run in two distinct environments:

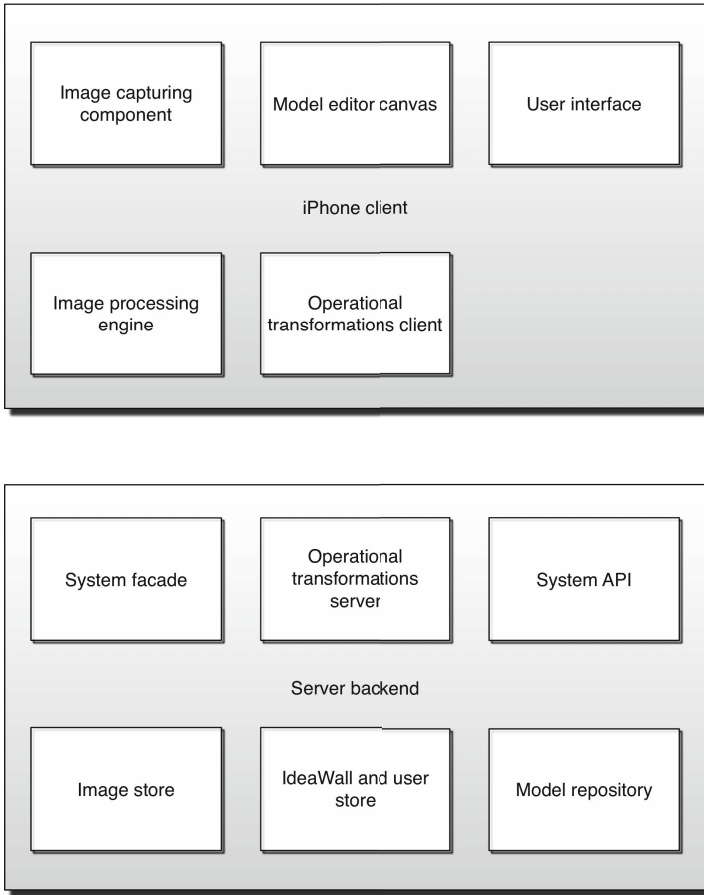
1. a "physical" location, usually a room, including:
  - a whiteboard / packing paper wall;
  - a mobile phone with a high resolution camera;
  - a projector (standard multimedia projector);
  - people collaborating within this room using post-it notes or cardboard notes on the whiteboard to brainstorm ideas, pens and erasers
2. and a "virtual" location, usually a client computer (including mobile devices) with an Internet browser and an Internet connection.

With this in mind, Figure 3 depicts a high level architecture of the solution. For simplicity, only a situation where a mobile phone is used in a "virtual" location is depicted. Majority of the client components is not required in a "virtual" location (effectively only a web browser is required), as compared to the "physical" location.

The depicted components include:

- an image capturing component, responsible for capturing images of the whiteboard and performing any required transformations (cropping, scaling, rotating, perspective correcting) in order to provide a normalized image of the whiteboard;
- model editor canvas, displayed through the projector on the whiteboard, so that the remote participants' contribution can be made visible, as well as providing feedback to the users (connection status etc.);
- user interface, providing access to all required functions, such as inviting new participants, creating new IdeaWalls etc.;
- image processing engine, performing object detection, using computer vision methods;
- operational transformations client, responsible for the client side functionality of the conflict resolution engine (if two participants perform conflicting actions on the same object), enabling smooth collaboration on the whiteboard, without a need to check out and check in parts of the model;
- system facade, providing access to system functions such as creating new users, authenticating, or creating new IdeaWalls, via a RESTful interface;
- operational transformations server, providing server-side conflict resolution methods;
- system API, together with system facade, providing complete programmatic access to the backend;
- image store, IdeaWall and user store, model repository, providing persistence layer for the system.





**Fig. 3.** High level architecture of IdeaWall

## 5 Related Work

### 5.1 User Interaction

The closest known related prior art is by Scott R. Klemmer and others [23,5]. Their work looks at integrating physical and digital interactions on whiteboards for fluid design collaboration. Klemmer and others have produced a prototype known as the Designers Outpost. The system is intended to support designers during the early stages of web site design. Like IdeaWall, the system enables designers to collaboratively author on a whiteboard augmented with cameras that capture documents placed on the wall.

However, Project IdeaWall holds certain differences and advantages from this prior art:

1. It does not offer support for multiple simultaneous users. The Designers Outpost allows co-existence of post-it notes and digital ink on a whiteboard, but does not support multiple users (i.e. multiple whiteboards condensed in one) simultaneously. On the other hand, IdeaWall is more advanced as it can support multiple users simultaneously.
2. It does not distinguish or identify more than one type of artefact on the whiteboard (i.e. it does not distinguish artefacts other than post-it notes, relationships and digital writing). On the other hand, IdeaWall can recognise more than ink and post-it notes i.e. it can distinguish BPM symbols, such as events and functions, in addition to relationships and digital writing.
3. It is primarily for web design purposes (e.g. web site architecture diagrams). IdeaWall, on the other hand, can be used for a number of use cases, including process modelling and requirements engineering. Also, whilst this work transferred whiteboard data into a web authoring tool, information on IdeaWall collaboration boards will be available in Web authoring tools straight away.

The approach of Klemmer et al. therefore has a web-focus, does not support multiple users simultaneously, and cannot distinguish/identify artefacts other than post-it notes, relationships and digital writing.

## 5.2 Operational Transformations

IdeaWall builds on top of the operational transformation (OT) work of Hettel and Balko. Operational transformations [10] ensure consistency of physical media (e.g. post-it notes) being exchanged in IdeaWall. Real-time collaborative editing of whiteboard information is possible by building on top of this work.

Other related topics include understanding of user awareness in such distributed collaboration scenarios [8], including proper design of such spaces [9].

## 6 Summary and Outlook

Project IdeaWall gives workers the ability to use traditional (physical) brainstorming tools to share information and collaborate with teams at other locations. Each person in the collaboration can be active (i.e. any participant can modify the whiteboard contents) and changes can be replicated simultaneously at all locations. Furthermore, IdeaWall:

1. supports the early stages of creative processes;
2. supports physical and virtual whiteboard collaboration between multiple users simultaneously;
3. enables the sharing of multiple different artefacts (post-it notes, whiteboard drawings, relationships, BPM notation, etc.) between remote and distributed teams;

4. transforms temporary/short-term whiteboard diagrams into long-term documentation (i.e. whiteboard diagrams can be transferred to PC tools for later use);
5. and generally improves collaboration between geographically dispersed teams.

We are currently focusing on extending the functionality of IdeaWall with numerous features requested by the early testers of the application. These include features such as gesture recognition, handwritten text recognition and ability to use more sophisticated whiteboard backgrounds (such as maps). In parallel, we are starting extensive user research, to identify the pain points of such multi modal distributed collaboration application on mobile devices.

**Acknowledgements.** This document is a report of work of a larger team of researchers and developers. While the architecture, research and work has been driven by the author, many others have contributed to the project. The author wishes to express his gratitude to those who have been contributing to the project.

## References

1. Rittenbruch, M., Kahler, H., Cremers, A.B.: Supporting collaboration in a virtual organization. In: Proceedings of the International Conference on Information Systems (ICIS 1998), pp. 30–38. Association for Information Systems, Atlanta (1998)
2. Klemmer, S.R., Newman, M.W., Farrell, R., Bilezikjian, M., Landay, J.A.: The designers' outpost: a tangible interface for collaborative web site. In: Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology (UIST 2001), pp. 1–10. ACM, New York (2001), <http://doi.acm.org/10.1145/502348.502350>, doi:10.1145/502348.502350
3. Everitt, K.M., Klemmer, S.R., Lee, R., Landay, J.A.: Two worlds apart: bridging the gap between physical and virtual media for distributed design collaboration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003), pp. 553–560. ACM, New York (2003), <http://doi.acm.org/10.1145/642611.642707>, doi:10.1145/642611.642707
4. Klemmer, S.R.: Integrating Physical and Digital Interactions. *Computer* 38(10), 111–113 (2005), <http://dx.doi.org/10.1109/MC.2005.343>, doi:10.1109/MC.2005.343
5. Klemmer, S., Everitt, K., Landay, J.: Integrating Physical and Digital Interactions on Walls for Fluid Design Collaboration. *Human-Computer Interaction* 23(2), 138–213 (2008), <http://www.informaworld.com/openurl?genre=article&doi=10.1080/07370020802016399&magic=crossref>
6. Yang, J., Papazoglou, M.P.: Interoperation support for electronic business. *Commun. ACM* 43(6), 39–47 (2000), <http://doi.acm.org/10.1145/336460.336473>, doi:10.1145/336460.336473
7. Cherubini, M., Venolia, G., DeLine, R., Ko, A.J.: Let's go to the whiteboard: how and why software developers use drawings. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007), pp. 557–566. ACM, New York (2007), <http://doi.acm.org/10.1145/1240624.1240714>, doi:10.1145/1240624.1240714

8. Mcewan, G., Rittenbruch, M., Mansfield, T.: Understanding awareness in mixed presence collaboration. In: Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI 2007), pp. 171–174. ACM, New York (2007), <http://doi.acm.org/10.1145/1324892.1324924>, doi:10.1145/1324892.1324924
9. Broughton, M., Paay, J., Kjeldskov, J., O'Hara, K., Li, J., Phillips, M., Rittenbruch, M.: Being here: designing for distributed hands-on collaboration in blended interaction spaces. In: Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7 (OZCHI 2009), pp. 73–80. ACM, New York (2009), <http://doi.acm.org/10.1145/1738826.1738839>, doi:10.1145/1738826.1738839
10. Sun, C., Ellis, C.: Operational transformation in real-time group editors: issues, algorithms, and achievements. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work (CSCW 1998), pp. 59–68. ACM, New York (1998), <http://doi.acm.org/10.1145/289444.289469>, doi:10.1145/289444.289469

# A Novel Approach to Increase Efficiency of OSS/BSS Workflow Planning and Design

Tetiana Kot<sup>1</sup>, Andrey Reverchuk<sup>3</sup>, Larysa Globa<sup>1</sup>, and Alexander Schill<sup>2</sup>

<sup>1</sup> National Technical University of Ukraine «Kyiv Polytechnic Institute», Ukraine

<sup>2</sup> Technische Universität Dresden, Fakultät Informatik, Deutschland

<sup>3</sup> SITRONICS Telecom Solutions, Czech Republic a.s.

**Abstract.** Nowadays, communication technologies and the range of mobile operator services are changing extremely fast. This results in the need for constant adaptation and expansion of OSS/BSS<sup>1</sup> used in mobile operator networks. Currently, adaptation and expansion strategies are poorly formalized and validated. In current state-of-the-art approaches, several iterations involving analysts and system architects are necessary, resulting in time and money consuming service development. The workflow design method proposed in this paper fills this gap. It employs a well-defined workflow and analysis model for developing and adapting OSS/BSS. The applicability of this novel approach is confirmed by a prototypically implemented design software tool which has been tested in a telecommunication enterprise. The developed tool provides automation of service planning, computational independent workflow design and its transformation into its realization model. The reduction of development time and thus necessary financial input has been proven based on our real-world experiments.

**Keywords:** OSS/BSS, service planning, workflow design, Dia editor.

## 1 Introduction

*“The technology sector in general and the telecommunication companies in particular operate in increasingly competitive environments. The companies that survive and excel are those offering the most compelling range of products and services. Since the underlying technologies tend to offer similar features and functionalities, the only differentiation is the services created from these technologies. The method used to create a great service is service design.*

*Even to the professionals within the telecommunication and service provider sectors (i.e. the companies that provide telecommunication/Internet services), it is often difficult to articulate the concept and purpose of service design. However, I have seen so many projects and service developments fail because there were no service designers in the project team.”[1]*

---

<sup>1</sup> Operation Support System/Business Support System.

Telecommunication companies are notable for technological effectiveness and innovation, constantly renovating their technologies and services by adopting OSS/BSS - complex software systems, providing telecommunication companies functioning) [2]. OSS/BSS adaptation and expansion is realized via service planning, workflow design, realization and re-engineering according to the changes of requirements to services and OSS/BSS features. Additionally, mobile operators tend to minimize time of service provision, using constantly improving communication technologies and software applications. Thus, mobile operators require means and tools for fast workflow design and reengineering during OSS/BSS runtime and for providing services within a minimal time.

OSS/BSS adaptation and expansion mainly concerns the development of service provisioning applications and includes:

1. service planning, defining service provision time and resources;
2. computational independent workflow design, defining the order of tasks while service provisioning;
3. computing workflow design, enactment, monitoring and optimization.

Currently, the service planning stage is performed manually. This does not result in good solutions because a lot of factors have to be considered for finding the optimal solution when planning differentiated services [3]. Furthermore, in the current state-of-the-art computational independent workflow design, being performed, using existing notations and tools on one hand does not consider all required parameters, necessary on the service planning stage, such as numerical values of execution time and resources, required for that, and also document and information flows, supporting service provision, and on the other hand provides poor connection to system functionality, which should realize these workflows.

This paper describes a novel computational independent workflow design method, focusing on the workflow model and its analysis, allowing to automate service planning stages and to reduce the time and costs for OSS/BSS adaptation in general.

The paper is structured as follows: Section 2 contains state of the art analysis of workflow design notations and tools. Furthermore, service implementation technologies are described. Section 3 introduces workflow design methods, focusing on two core aspects: A computational independent workflow model, describing service provision and considering service provisioning parameters, required for its analysis, and a workflow analysis method, providing service provisioning time minimization. Section 4 presents a prototypical realization of the design method and highlights evaluation results of the developed tool. The evaluation has been applied using a real-world scenario within a telecommunication company. Section 5 concludes the work with a summary and outlook on future work.

## 2 State of the Art and Background

Service planning is defined by finding a good way to create service provision at minimal time, having specific resource values as a limitation. In the following a summary of the state of the art of the central areas of this overall field will be discussed.

A state of the art analysis of workflow design notations and tools are presented in section 2.1, workflow analysis methods and systems in section 2.2. Parameters, which should be taken into account at the planning stage are presented in section 2.3.

## 2.1 Computational Independent Workflow Design

Computational independent workflows are designed using graphical standards and allowing their formalization and their possible flows and transitions in a diagrammatic way. Analysis has shown that in practice computational independent workflows are usually designed using graphical notations such as BPMN 2.0<sup>2</sup>, UML AD<sup>3</sup>, USLD<sup>4</sup> and tools such as CA ERwin Process Modeler<sup>5</sup>, Enterprise Architect<sup>6</sup> and MS Visio<sup>7</sup>.

USDL is sufficiently generalized, still under development stage and doesn't fully meet all the requirements of OSS/BSS workflows analysis and design. Additionally, its usage is difficult due to its complexity, in spite of its comprehensiveness.

BPMN 2.0 shortcomings are clearly described in [4]. The central argument against using regular BPMN is that management of resources can be expressed only via lanes (actors, roles, etc.) or performers of user or manual tasks. No execution time parameters are considered. All further existing workflow modeling notations have this core criticism in common.

Nevertheless, BPMN, providing the ability of computational independent to computing workflows transformation (BPEL<sup>8</sup> diagrams), being widespread in industry, can be applied as a basic notation for OSS/BSS computational independent workflow design. Thus, it is intended to extend it adding the missing concepts.

## 2.2 Workflow Analysis

The given short overview of the workflow analysis methods and tools has shown, that there exist two types of analysis, both considering computational workflow:

1. Design time analysis (simulation and verification). Monte Carlo simulations can be used as well as Petri Nets analysis as there exists transformation approaches for BPMN [5], UML AD [6], EPC [7], BPEL [8] to Petri Nets with their further analysis. USLD diagrams can be analyzed using service ontological analysis [9].
2. Runtime analysis (for instance, process mining, based on the execution logs) [10].

Software tools such as Pegasus, Cactus, ASKALON, GLUE, etc. [11] are used for these analysis fields. All mentioned and analyzed current possibilities for this task stage are very limited. There are no tools available to automate the service planning

---

<sup>2</sup> <http://www.omg.org/spec/BPMN/2.0/>

<sup>3</sup> UML Activity Diagram.

<sup>4</sup> <http://www.internet-of-services.com/index.php?id=264>

<sup>5</sup> [http://erwin.com/products/detail/ca\\_erwin\\_process\\_modeler/](http://erwin.com/products/detail/ca_erwin_process_modeler/)

<sup>6</sup> <http://www.sparxsystems.com/products/index.html>

<sup>7</sup> <http://office.microsoft.com/en-us/visio/>

<sup>8</sup> Business Process Execution Language.

stage considering service provision time and resources. As an improvement, a graph model can be used to verify workflow diagrams connectivity, as presented in [12]. By this, a service provisioning time minimization can be realized. The methods, usually used on this stage, such as shortest path problem for a graph, are not applicable, due to the fact that workflows, describing service provisioning, contain parallel tasks, and each of them can have a few implementation variants. Thus, some other mathematical approach should be applied for this problem.

Lacks of existing tools and methods of workflow design and analysis make their usage while service planning and OSS/BSS workflow design complicated. The central criticism is that the requirements analysis stage is applied mainly in a manual manner.

### 2.3 Service Planning

In differentiated service models, used today by mobile operators, QoS<sup>9</sup> is realized by marking packets based in the customers' service class. In response to these markings, routers and switches use various queuing strategies to tailor performance to requirements. According to this, service provision time and resource vary potentially heavily.

Service provision, depending on content and communication technologies, is characterized by parameters such as QoS, acceptable service delay, capacity, depending on radio-technology (GPRS/EDGE<sup>10</sup>; LTE<sup>11</sup>), etc. Besides, separate tasks of the workflow can have a few variants of realization. For instance, subscriber requests can be transferred using technologies such as GSM<sup>12</sup>, GPRS, CDMA<sup>13</sup>; tariffing can be realized using internal online raters, internal offline raters or external raters. Their values define service provision time and resources, which are to be defined on the planning stage.

Thus, workflow tasks can be implemented in four different ways. The implementation variant is defined by time and resources, including those required for service provision, and total resource, required for service provision, is a sum of resources, required to implement each of the tasks of the workflow. It is defined by:

$$R = \sum_{k=1}^K r_{ki}^n \tag{1}$$

$n = \overline{1, N}$ , where N – is the number of workflow tasks, describing service provision;  $k = \overline{1, K}$ ,  $K \leq 5$ , where K is the number of implementation variants for the workflow tasks.

Thus, when applying service planning, operators work with the following data:

- R – total amount of resources required for service provision (workflow execution);

---

<sup>9</sup> Quality of Service.

<sup>10</sup> General Packet Radio Service/Enhanced Data rates for GSM Evolution.

<sup>11</sup> 3GPP Long Term Evolution.

<sup>12</sup> Global System for Mobile Communications.

<sup>13</sup> Code Division Multiple Access.



- $r_{kl}^n$  - resource, required for implementing task  $l$  of stage  $k$  using realization variant  $n$ ;
- $\xi_{kl}^2$  - execution time of task  $l$  of stage  $k$  using realization variant  $n$ .

For all task implementation variants, time dependence on the resource  $\xi(r)$  is not linear. This should be considered at the planning stage during the computational independent workflow design.

### 3 Workflow Design Method

According to [13], workflow design includes workflow modeling and simulation. The novel method of workflow design presented in this section is focused on computational independent workflow and consists of the following stages:

1. extended modeling of computational independent workflow;
2. computational independent workflow analysis, including:
  - forming workflow graph and verifying its connectivity;
  - workflow execution time minimization;
  - transformation of workflow to realization diagram.

The suggested method modifies the MDA<sup>14</sup> approach on the business logic level (fig.1.).

#### 3.1 Workflow Model

The workflow model is one of the core aspects of the proposed method, allowing its formal description and thus its in-depth analysis and transformation to more fine-grained representations. In the following the workflow formalization variant used within our engineering approach is presented.

The mathematical formalization of a workflow can be done by using:

$$BP = (E, I, P) \quad (2)$$

where  $E$  is the set of workflow identification objects;  $I$  is the set of workflow informational objects and  $P$  is the set of workflow parameters, characterizing service provisioning.

The identification objects  $\{E_{id}, id=1,4\}$  include:  $E_1$  = name,  $E_2$  = description,  $E_3$  = executor,  $E_4$  =  $O$  – set of works.

The set of workflow informational objects includes income and outcome document and data objects:

$$I = \{I_{doc}^{in}\} \cup \{I_{dat}^{in}\} \cup \{I_{doc}^{out}\} \cup \{I_{dat}^{out}\} \quad (3)$$

The workflow parameters  $\{P_i, i=1,6\}$  cover:  $P_1 = T_{ex}$  – workflow execution time;  $P_2 = R$  – resource, required for workflow execution;  $P_3 = A$  – ability to be automatically executed;  $P_4 = S$  – set of OSS/BSS subsystems, used for workflow execution;  $P_5 = F^S$  – set of OSS/BSS separate subsystem functions, realizing task execution;  $P_6 = P_{ad}$  – set of additional workflow parameters.

---

<sup>14</sup> Model Driven Architecture.

Furthermore, separate work models can be represented in a formal manner as:

$$O = (O, I^O, P) \tag{4}$$

where  $O$  is the set of identification objects;  $I^O$  is the set of informational objects;  $P$  is the set of parameters, characterizing service provisioning concerning separate work realization.

The set of identification objects  $\{O_{id, id=1,3}\}$  includes:  $O_1 = N_o$  - name;  $O_2 = d$  - description;  $O_3 = E$  - executor. The set of task informational objects includes income and outcome informational objects, including document and data objects. Set of work parameters  $\{P^O_{i, i=1,7}\}$  include:  $P^O_1 = \xi_{kl}(r_{kl})$  — execution time of work  $l$  of stage  $k$ ;  $P^O_2 = r_{kl}$  - resource, required for execution work  $l$  of stage  $k$ ;  $P^O_3 = a$  - the ability to be automatically executed;  $P^O_4 = S$  - set of OSS/BSS subsystems, used for workflow execution;  $P^O_5 = F^S$  - set of OSS/BSS separate subsystem functions, realizing works execution;  $P^O_6 = R_O^n$  - work realization alternatives, defining execution time and resource values:

$$R_O^n = (N_R, \xi_{kl}^n(r_{kl}), r_{kl}^n) \tag{5}$$

$P^O_7 = P_{ad}$  - set of additional work parameters.

In more details the presented model is described in [14]. It will allow to perform workflow analysis at a planning stage, applying graph theory and optimization algorithm, represented below.

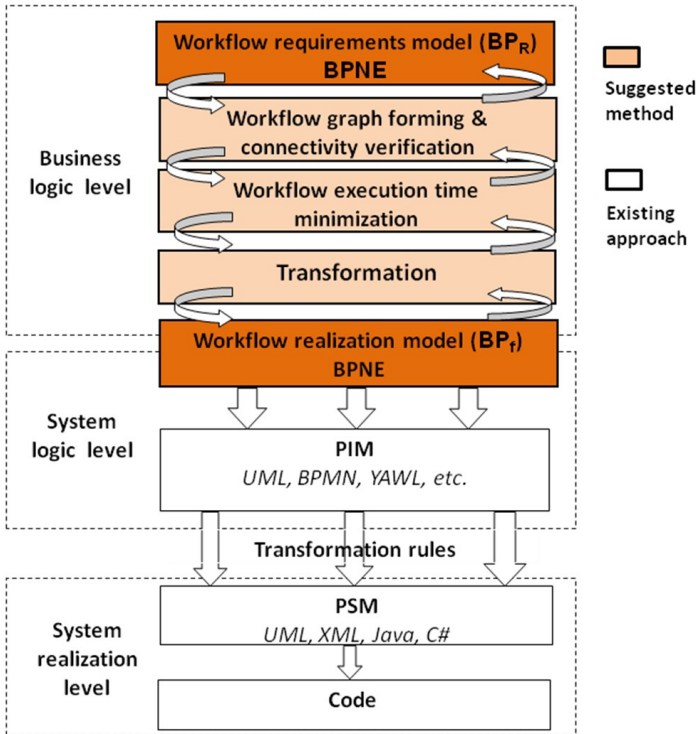


Fig. 1. Workflow design method extending the OMG MDA approach

### 3.2 Workflow Analysis Method

The workflow analysis method is the second central point of the proposed design method.

The workflow graph model can be represented as sequential stages, containing a few parallel executed tasks or just one task (fig. 2), which enables the definition of workflow execution time as follows:

$$T_{ex} = \sum f_{st} = \sum_k \max_l \xi_{kl}(r_{kl}); \tag{6}$$

The suggested method of workflow analysis, providing workflow model verification, execution time minimization and automating its transformation, can be represented mathematically as follows:

$$M = (G_f, G_v, M_{min}, M_{tr}) \tag{7}$$

where  $G_f$  is the graph generating procedure;  $G_v$  is the graph connectivity verification;  $M_{min}$  is the execution time minimization and  $M_{tr}$  is the diagram representing the model to realization transformation. According to the proposed method, workflow analysis is performed as follows. When the workflow diagram is designed, having a few implementation variants of its operations, the workflow graph is generated and its connectivity and syntax are verified. The execution time of the verified workflow model can be found in the next step, having the general resource value as limitation. After the time minimization procedure is realized, only one implementation variant for each operation is selected. The workflow model with minimal execution time is transformed into the realization model, describing the system modules, their functions, their time and resources requirements.

**Generating the Workflow Graph and Its Connectivity Verification.** The described workflow model allows requesting the control graph and information flow graph. Graph connectivity is verified using standard procedures and algorithms, described in.

**Workflow Execution Time Minimization.** Each task of the workflow has one to three implementation variants, defining execution time and resources, the task is to find such implementation variant for each task to minimize the total execution time of the workflow (6), when the total resource (1) is limited and known.

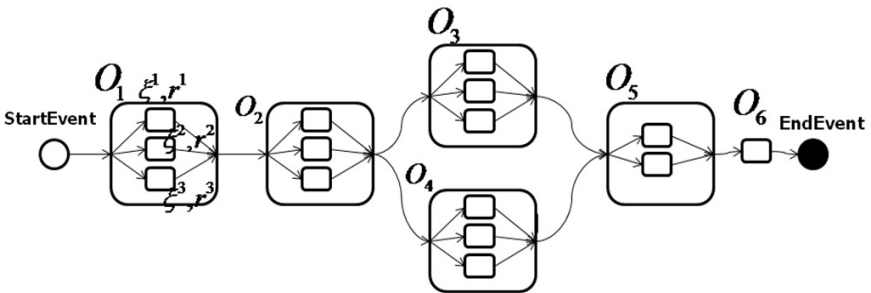


Fig. 2. Workflow graph with task implementation alternatives

According to practical experience, the number of typical software modules providing one service (implementing one workflow) is not more than 10. Thus, the number of parallel executed tasks can't exceed 10 and usually is not more than 5. Then the task of finding minimal workflow execution time is P-complete and can be solved in polynomial time.

The objective function of the task is represented in (8)

$$F(r) = \min_{\sum_{kl} r_{kl} = r} M \sum_k \max_l \xi_{kl}(r_{kl}) \tag{8}$$

Furthermore, it is necessary to find the implementation variant for each task, i.e.  $\{r_{kl}, k = 1, \dots, n, l = 1, \dots, m\}$ , where the required minimum of time is reached (8).

For finding the function specified in (8), dynamic programming is applied. This is realized in a two steps approach (fig.3):

1. determining minimal time of tasks executed in parallel works;
2. determining minimal time of workflow sequential stages, using results from the previous step.

Thus, using dynamic programming algorithms, the task of workflow analysis on the service planning and design stage, can be solved.

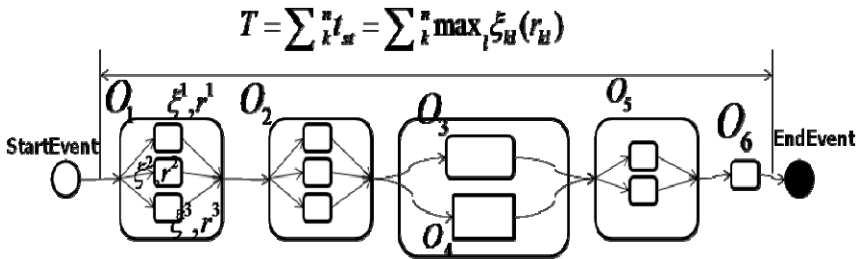


Fig. 3. Workflow time minimization

**Transformation of Workflow Requirements into a Realization Diagram.** The approach of transforming the workflow requirements into a realization diagram includes three core steps:

- 1) defining the ability of work to be automatically executed:  $P^O_3 = a$  [yes/no/half];
- 2) generating the workflow intermediate model:
  - adding each of the task parameters – OSS/BSS subsystems and their functions;
  - analyzing task execution time and resources: dividing them in functional and non-functional components;

3) generating a workflow realization diagram, containing only functional components of time and resources, system modules and functions description.

The workflow realization diagram describes requirements to workflow realization and can be used for architecting the service provisioning application design.

## 4 BPMA

The proposed method has been implemented based on the workflow design tool BPMA<sup>15</sup>. It is realized using GTK+, the Dia diagram editor, PyDia interface, Python Interpreter, PyGTK<sup>16</sup> and BPEA – a module for setting workflow parameters and analyzer. BPEA is a core BPMA component, implementing suggested workflow modeling and its analysis algorithms. Its functional scheme is represented in fig. 4. BPEA includes five main submodules: “init”, “props”, “bplyzer”, “transform” and “reports”. Submodule “init” realizes Dia and user intercommunication. It provides an user interface, checks user commands conformance and data correctness, and also launches all module functions. The submodule “props” provides setting, changing and saving of workflows and its objects parameters. “Bplyzer” implements time minimization algorithms. The submodule “transform” implements the transformation logic to create a realization diagram from workflow requirements. The submodule “reports” generates and represents reports regarding workflow modeling and analysis results.

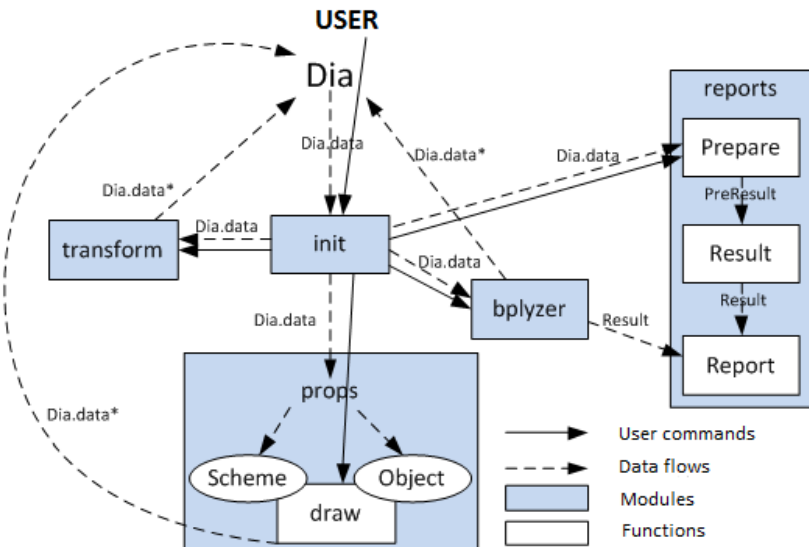


Fig. 4. BPMA functioning scheme

<sup>15</sup> Business Process Modeling & Analysis.

<sup>16</sup> Set of Python wrappers for the GTK+ GUI library.

Figure 5 visualizes a workflow developed with the created tool.

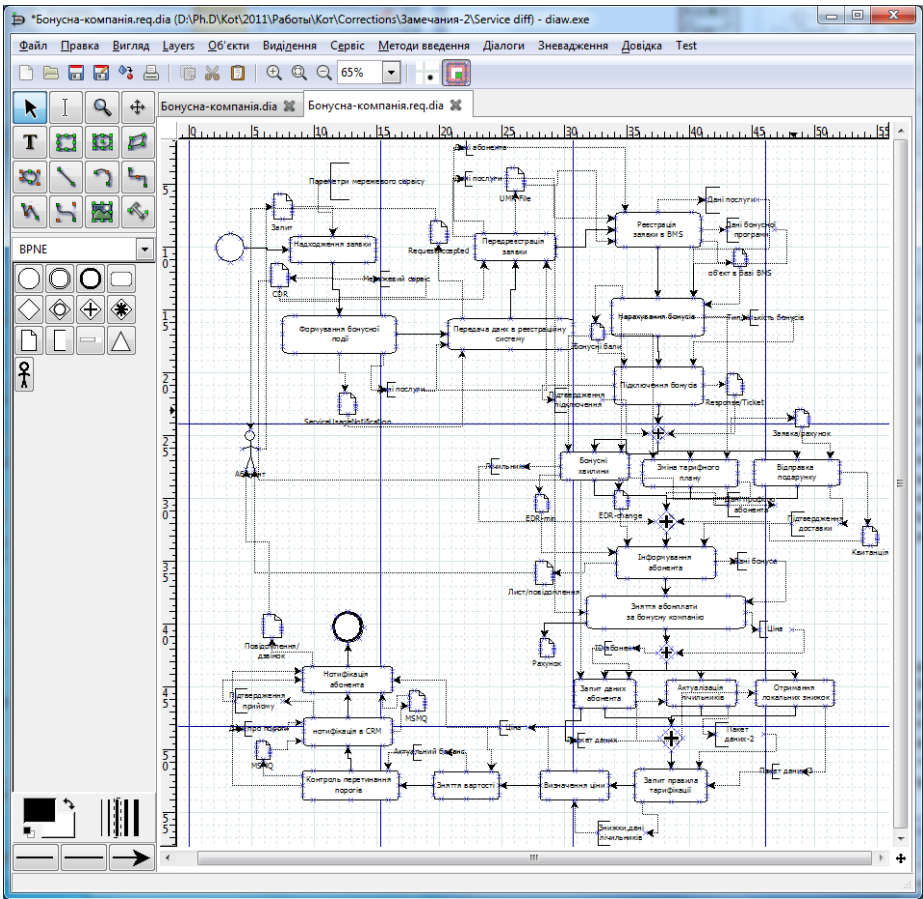


Fig. 5. Example workflow used within a case study

BPMA has been tested intensively during the planning and design of several services in the SITRONICS Telecom Solutions company. Testing results have proven its ability to reduce time and costs of the service planning stage and service development in general in comparison to existing tools, used when service design (BPWin, Enterprise Architect, etc.). Thus, for services, having 1-1.5 months of development time, it can be reduced by 3-5 days and development costs can be decreased by 5-7 man-days for one service. The proposed workflow analysis can reduce the time used for service provision up to 20 seconds for the services, provided in 3-5 minutes by finding the combination of tasks implementation variants, having the same resource limit.

## 5 Summary and Outlook

A novel approach for analyzing and developing OSS/BSS workflows has been proposed in this paper. The discussion focused on the development of a computational

independent workflow model and the procedure of its analysis. The proposed model provides formalization of service provision parameters, required for the planning stage. The analysis methods allow automating service planning and minimizing service provision time. The OSS/BSS workflow design method reduces the time of OSS/BSS adaptation up to 10% by improving and automating the service planning stage.

Future work will focus on transforming the proposed workflow model to an executable one, providing required parameters transfer, its enactment, monitoring and optimization both on the design and on runtime stages.

## References

1. Pang, S.: Successful service design for telecommunications: a comprehensive guide to design and implementation, p. 351. John Wiley & Sons Ltd. (2009)
2. Terplan, K.: OSS Essentials: Support System Solutions for Service Providers / Kornel Terplan, p. 610. John Wiley, New York (2001)
3. RFC 2474: Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, <http://tools.ietf.org/html/rfc2474>
4. Börger, E.: Approaches to modeling business processes: a critical analysis of BPMN, workflow patterns and YAWL, p. 14. Springer (September 2011)
5. Transformation of BPMN models for Behaviour Analysis, [http://www.win.tue.nl/~jmw/\\_media/public/transformationforbehaviouranalysis.pdf](http://www.win.tue.nl/~jmw/_media/public/transformationforbehaviouranalysis.pdf)
6. Baresi, L., Pezzè, M.: On Formalizing UML with High-Level Petri Nets. In: Agha, G., De Cindio, F., Rozenberg, G. (eds.) APN 2001. LNCS, vol. 2001, pp. 276–304. Springer, Heidelberg (2001)
7. Nüttgens, M., Feld, T., Zimmermann, V.: Business Process Modeling with EPC and UML: Transformation or Integration? In: Schader, M., Korthaus, A. (Hrsg.) Proceedings of the Unified Modeling Language - Technical Aspects and Applications, Mannheim, Heidelberg, October 1997, pp. S250–S261 (1998)
8. Ouyang, C., Verbeek, E., van der Aalst, W.M.P., Breutel, S., Dumas, M., ter Hofstede, A.H.: Formal Semantics and Analysis of Control Flow in WS-BPEL. Technical report (revised version), Queensland University of Technology (October 2005)
9. Oberle, D., Bhatti, N., Brockmans, S., Niemann, M., Janiesch, C.: Countering Service Information Challenges in the Internet of Services. *Journal of Business & Information System Engineering* 1(5), 370–390 (2009)
10. Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M.: Workflows for e-Science, Scientific Workflows for Grids, 1st edn., p. 552. Springer (2006)
11. van der Aalst, W.: BPM and Workflow Analysis. *BPTrends* 5(4), 1–2 (2007)
12. Globa, L., Kot, T., Schill, A., Strunk, A.: Method of IBIS design and workflow realization. *Polish J. of Environ. Stud.* 18(4a), 35–38 (2009)
13. Ko, R.K.L., Lee, S.S.G., Lee, E.W.: Business Process Management (BPM) Standards: A Survey. *Business Process Management Journal* 15(5), 48 (2009)
14. Kot, T., Globa, L., Schill, A.: Applying business process modeling method when telecommunication services development. In: Proceedings of 21st International Crimean Conference Microwave and Telecommunication Technology, CriMiCo 2011, Sevastopol, Ukraine, vol. 1, pp. 457–458 (2011)

# Recognition and Pseudonymization of Personal Data in Paper-Based Health Records

Stefan Fenz<sup>1</sup>, Johannes Heurix<sup>2</sup>, and Thomas Neubauer<sup>1</sup>

<sup>1</sup> Vienna University of Technology, Vienna, Austria  
{stefan.fenz, thomas.neubauer}@tuwien.ac.at

<sup>2</sup> SBA Research, Vienna, Austria  
jheurix@sba-research.org

**Abstract.** E-health requires the sharing of patient-related data when and where necessary. Electronic health records (EHR) allow the structured and expandable collection of medical data needed for clinical research studies and thereby not only enable the optimization of clinical studies, but also results in higher statistical significance due to a larger number of samples. While the digitization of medical data and the organization of this data within EHRs have been introduced in some areas, massive amounts of paper-based health records are still produced on a daily basis. This data has to be stored for decades due to legal reasons but is of no benefit for research organizations, as the unstructured medical data in paper-based health records cannot be efficiently used for clinical studies. Furthermore, legal regulations prohibit the use of documents containing both personal and medical data for clinical studies, which leads to expensive data acquisition phases and limited samples. This paper presents the MEDSEC system for the recognition and pseudonymization of personal data in paper-based health records. MEDSEC integrates unique methods for (i) automatically identifying personal and medical data, (ii) automatically annotating the optical character recognition (OCR) output data of paper-based health records with standard-compliant metadata, and (iii) automatically pseudonymizing the personal data. With MEDSEC, health care organizations profit by (i) strengthening clinical research resulting in faster and more reliable results and reduced costs, and (ii) providing an environment of trust for its patients and employees that guarantees privacy.

**Keywords:** EHR, privacy, annotation, HL7 CDA, pseudonymization, transformation, OCR.

## 1 Introduction

In today's health care system, the availability of sound information has tremendous impact on decisions regarding patients' care and, as a result, on the quality of treatment and patients' health. The digitization of medical data (e.g., by using electronic health records (EHR)) promises (i) the reduction of adverse drug events accounting for about US\$175 billion a year in the US, (ii) the reduction of the very high number of more than 200,000 cases of deaths a year in the US [1] as it provides physicians and their health care teams [2] with decision support systems and guidelines for drug interactions, and (iii) massive savings that can be achieved by digitizing diagnostic tests and images.



A study by the non-profit research organization Rand Corporation found out that adopting the EHR could result in more than US\$81 billion in annual savings in the US if 90% of health care providers used it [1].

In addition to the direct benefits, the digital storage and analysis of medical data could mean a quantum leap in clinical research, because it allows the improvement of communication between health care providers and of access to data and documentation, leading to better clinical and service quality [3]. Today, the success of clinical trials heavily depends on the recruitment of enough eligible participants in a timely manner. Failing to meet recruitment goals can hamper the development and evaluation of new therapies and can not only increase drug development costs but also health care system costs (cf. [4] for estimates about the costs of clinical trials). Today, 86% of all trials fail to start on time because subjects cannot be recruited in time and because only 7% of eligible patients enroll in a clinical trial. One study, which looked at 4,000 clinical trials over five years, discovered that nearly half of the time spent on the trial process involved patient, site and investigator recruitment [5]. Clinical research is ending up in a vicious circle because clinical trial capacity does not meet the demand, and whereas the number and the duration of trials is increasing, the number of patients available for trials is decreasing. The structured organization of digitized medical data (e.g., within an EHR) allows (i) the more effective and efficient recruitment of clinical trial participants, (ii) the reduction of administrative overhead, (iii) the impact reduction of data errors due to larger samples and (iv) the faster identification of adverse outcomes. However, the vast majority of health records is still only available on paper and experts agree that the amount of paper-based health records will never be beat down below 20%, leaving enormous potential for improving clinical research. There are three major problems preventing the use of paper-based health records in clinical research:

- First, paper-based health records do not provide machine-interpretable metadata and circumvent the automatic identification of personal and medical data elements. Currently, no methods for the automatic identification of personal and medical data exist. Existing high-level privacy taxonomies (e.g., [6]), ontology-based trust negotiation approaches (e.g., [7]), and web standards, such as W3C P3P [8], provide a categorization of privacy-relevant data items but do not provide common synonyms and formal specifications of personal and medical data elements to enable their automatic detection in paper-based health records. Since it is not possible to use only the content of a data element to automatically determine its type, formal descriptions have to include potential identifiers used in paper-based health records. The further use of digitized and pseudonymized paper-based health records for clinical research highly depends on the complete identification of personal and medical data.
- Second, the sole digitization of paper-based health records is not sufficient for providing clinical research with suitable data. In addition to the actual digitization, it is of paramount importance for the distinction between different data elements to enrich the gathered data with appropriate standard-compliant metadata (e.g.,

according to the HL7 standard). While products for indexing optical character recognition (OCR) output data by self-defined profiles exist [11], no methods are available for the automatic annotation of personal and medical data. The lack of such methods prevents the further processing of the gathered OCR output for clinical research. Existing methods for enriching OCR output with standard-compliant and appropriate metadata do not meet clinical research requirements and do not guarantee the complete identification of personal data according to the Austrian Data Protection Act. Furthermore, the data complexity in the health care domain and the need for exchanging this data over existing system boundaries requires the usage of standardized data structures and communication protocols. While standards such as HL7 are already implemented in several health care information systems, no methods for the automatic transformation of semantically enriched OCR output data into a standard such as HL7 exist.

- Third, privacy is one of the fundamental issues in health care today. With informative and interconnected health-related data comes highly sensitive and personal information. Due to the high sensitivity of the data, there is increasing social and political pressure to prevent the misuse of personal health data. It is the fundamental right of every citizen to demand privacy (cf. HIPAA, EU Directives), and furthermore, the disclosure of medical data can cause serious problems for the patient. The increasing fear of data abuse as well as the adoption of laws lead to the development of a variety of techniques for protecting patients' identity and privacy. The concept of pseudonymization (cf. [9][10]) allows the data to be associated with a patient only under specified and controlled circumstances. Existing approaches can be differentiated into two groups: the first group of approaches has major security shortcomings (cf. [11][12][13][14][15][16]); the second group solves these shortcomings, but is not designed for the centralized (mass) pseudonymization of data (due to different requirements regarding architecture, security, and performance).

## 2 Background

The annotation of OCR output data with appropriate metadata (e.g., birth date, first name and gender) requires the formal specification of what personal and medical data actually is. The most mature approach for classifying personal data is the Platform for Privacy Preferences Project (P3P) [8]. P3P Specification 1.1 [8] defines a base data schema for personal data, including data elements such as first name, birth date, phone number, and email. The ICD-10 Standard is an international statistical classification of diseases and related health problems. Together with the HL7 standard it can be used as a foundation to classify medical-related data elements.

While the mentioned data schemes outline personal and medical data elements, they do not describe how concrete instances of these data elements could look like (e.g., that

<sup>1</sup> Dynamic Zone OCR: <http://www.simpleindex.com>

<sup>2</sup> docWorks: <http://www.content-conversion.com>

<sup>3</sup> ImageNet: <http://www.miteksystems.com>

a name does not include any numbers). Another shortcoming is that no synonyms are given for the different data elements (e.g., gender/sex, first name/given name). We use the HIPAA PHI schema and the ontology-based trust negotiation approach (cf. [7]) as the basis for the development of a personal data ontology that includes common synonyms in multiple languages and formal descriptions that enable the automatic identification of personal data elements in health records. On the medical side we will use the HL7 standard as the foundation for creating common multi-lingual synonyms and formal descriptions for relevant medical data elements.

Besides the mere identification of personal and medical data elements, it is crucial to annotate the identified data elements with metadata that corresponds to well-established health care standards (e.g., HL7). We plan to combine existing indexing tools with the developed formal data element descriptions to automatically annotate personal and medical data elements.

The use of open standards can considerably reduce the costs of electronic data capture in clinical research. The CDISC ODM<sup>4</sup> is oriented towards drug development and clinical research. CDISC, for example, allows the automatic setup of the EDC system, the creation and instantiation of the database, and the full automation of the creation of electronic case report forms. HL7 is a standards development organization dealing with data standards for all health care operations. Because of its broader scope, HL7 has not dealt much with the nuances of clinical trials, while CDISC has not dealt with health care applications important to HL7, such as reimbursements and order processing. The HL7 Clinical Document Architecture (CDA) is a document markup standard that specifies the structure and the semantics of clinical documents in Extensible Markup Language (XML). Persistence, stewardship, potential for authentication, wholeness and human readability are the main characteristics of the CDA [17]. One of its main characteristics, however, is also its main downside: Human readability allows the convenient use in health care environments but inhibits privacy.

In order to protect patients' privacy when using, transferring and storing medical records, a variety of privacy enhancing technologies (cf. [18]) have been proposed. However, existing approaches (i) do not comply with the current legal requirements (cf. [19][20][21][22][23]), (ii) do not fulfill basic security requirements (cf. [24][25]), and (iii) are not applicable for use with clinical studies. In 2006, the United States Department of Health & Human Services issued the Health Insurance Portability and Accountability Act (HIPAA) [26], which demands the protection of patients' data that is shared from its original source of collection. While no explicit European standards regarding the protection of PHI exist, HIPAA defines 17 PHI identifiers that have to be removed from the health record: (i) names, (ii) locations, (iii) dates, (iv) ages greater than 89, (v) telephone numbers, (vi) fax numbers, (vii) email addresses, (viii) social security numbers, (ix) medical record numbers, (x) health plan beneficiary numbers, (xi) account numbers, (xii) certificate numbers, (xiii) vehicle identifiers, (xiv) device identification numbers, (xv) URLs, (xvi) IP addresses, (xvii) biometric identifiers, and (xviii) any other unique identifying number, code, or characteristic such as full face photos.

---

<sup>4</sup> Clinical Data Interchange Standards Consortium: Specification for the Operational Data Model, <http://www.cdisc.org>

Since 2005, the processing and movement of personal data in the EU has been legally regulated by Directive 95/46/EC [19]. A citizen's right to privacy is also recognized in Article 8 [27] of the European Convention for the Protection of Human Rights and Fundamental Freedoms. Additionally, domestic acts in many EU member states contain strict regulations for the processing of personal data.

Please note that e-health and especially clinical studies demand the pseudonymization of data: (i) Anonymization - the removal of the identifier from the medical data - has the major drawback that patients cannot profit from the results gained in the clinical studies (e.g., patients cannot be informed about actual findings such as newly developed medical treatments or major changes in the healing progress). (ii) Encryption assures patients' privacy by encrypting the medical records with the patients' private key. However, encrypted data cannot be used for clinical research (secondary use) without the explicit permission of the patient who has to decrypt the data and in doing so, reveals her identity. Pseudonymization is a technique where identification data is transformed and then replaced by a specifier that cannot be associated with the identification data without knowing a certain secret [9,25,10]. Pseudonymization allows the data to be associated with a patient only under specified and controlled circumstances. A pseudonymized database must contain at least two tables, one where all the personal information is permanently stored, and one where the pseudonyms and the pseudonymized data are stored. The process of identifying and separating personal from other data is called depersonalization. After depersonalization and subsequent pseudonymization, a direct association between individuals and their data cannot be established. However, existing approaches and systems have a variety of shortcomings. The system developed by Thielscher et al. (cf. [12]) relies on a centralized patient pseudonym list which provides a fallback mechanism in case a patient loses her smart card, as otherwise there would be no way to recover the identifier. Thielscher et al. circumvent the security flaw of a centralized patient pseudonym list by operating it off-line. This organizational work-around seems to promise a higher level of security until a social engineering attack is conducted on a person inside the system [28,29] or an attacker gains physical access to the computer that holds the list. The approaches developed by Pommerening (cf. [15,16]) use a combination of a hashing and an encryption technique. The encryption itself is based on a centralized secret key, which opens a vulnerability, as an attacker who knows this single key might gain access to all patients' medical data. The approach developed by Peterson [11] comes with some serious drawbacks: As all keys needed for decrypting the medical data are stored in the database, an attacker gaining access to the database could decrypt all information. Even more importantly, as the password is also stored in the database as well as the keys, the attacker could change data stored in the database. The architectures proposed by Schmidt et al. [30] and the Fraunhofer Institute, supported by the German Federal Ministry of Health [31,32], are based on encryption. As a result the data is fully encrypted, which is not practicable for the use in clinical studies.

The PIPE framework is a new patented architecture (cf. [33,34,35,36,37,38,39,40] for details on our patent and previous work) that improves existing approaches by (i) allowing the authorization of health care providers or relatives of the patient to access

specified medical data at encryption level, (ii) providing a secure fallback mechanism in case the security token is lost or worn out, (iii) storing the data without the possibility of data profiling, and (iv) allowing secondary use without establishing a link between the data and the person it refers to. Patient-identifying details are separated from the actual health data, resulting in detached data records. The relation between the patient and her health data is established with pseudonyms that are accessible only under specifically defined conditions. In this way, only persons who know the pseudonyms are able to link the patient with the health data. Pseudonyms are also used for data access permissions, e.g., defining new pseudonyms for access authorizations or revoking access rights by deleting the pseudonyms.

Apart from the security shortcomings, existing pseudonymization approaches - including the PIPE approach - have a number of characteristics in common:

- They depend on the smart card's crypto chip for performing cryptographic operations. Although this technique, combined with a certified card reader and a PIN, can be considered secure [41], it is not usable if central and automatic pseudonymization (e.g., in the case of pseudonymizing large amounts of data) is needed.
- They do not provide high performance (e.g., 12 millions documents a year) solutions for central pseudonymization. The cryptographic chip on the smart card does not provide anything close to the performance needed for pseudonymizing such a number of documents.
- There is no access to the data owner's card at the moment of pseudonymization. It would be logistically impossible to gain synchronous access to the data owner's keys. Asynchronous options are not considered in current architectures.
- The architectures are designed for patient-centric scenarios (e.g., use in EHRs) but not for allowing central pseudonymization while at the same time guaranteeing a high level of security and privacy.

What is required for the pseudonymization of data archives is a (i) central, (ii) high-performant, and (iii) automatic pseudonymization approach. In this proposal we define such an approach as 'mass pseudonymization'.

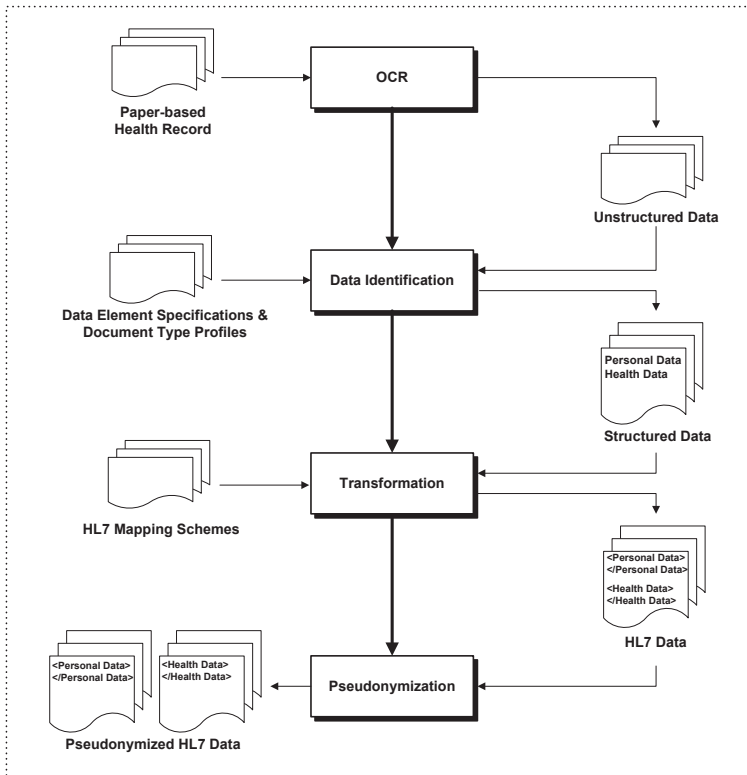
### 3 The MEDSEC System

The goal is to provide clinical studies with pseudonymized and structured medical data gained from existing paper-based health records. The proposed technical solution is divided into four main phases. Figure 1 shows an overview of the proposed solution.

*OCR:* The purpose of this phase is to digitize the content of paper-based health records. As the development of OCR engines was not the focus of this project, we use Google's open-source OCR engine Tesseract<sup>5</sup>, which is one of the most accurate open source OCR engines available<sup>6</sup>. Besides digitizing the actual content of paper-based health records, we enrich the corresponding OCR output with metadata containing information

<sup>5</sup> Tesseract: <http://code.google.com/p/tesseract-ocr/>

<sup>6</sup> Willis, Nathan (2006). Google's Tesseract OCR engine is a quantum leap forward: <http://www.linux.com/articles/57222>



**Fig. 1.** System overview

about the document type (e.g., physician’s letter, medical evidence, etc.). In our case we are using separate sheets with bar codes to identify the document type of each health record. As the assignment of these description sheets is a cumbersome manual process, we use a method for recognizing the document type automatically in the OCR phase. By matching the gathered layout data with the developed layout profiles we can improve the efficiency of the document type classification.

*Data Identification:* The data identification phase transforms the unstructured OCR data into a structured data format using the developed document type profiles. The document type profiles provide offsets for each data element/document type combination and enable us to identify data elements based on their position on the original health record. Additionally, we developed methods for identifying personal and medical data independently of an existing document type classification:

- Content-based identification: the data element is identified based on the content of the data item in question, e.g., checking each 10-digit number for its potential to be a social security number by calculating the check-digit or matching a string against a list of given names. We use the HIPAA PHI schema as the basis for categorizing privacy-relevant data elements and the HL7 data classification for categorizing

medical data elements. Based on that categorization we define synonyms and formal specifications of personal and medical data elements to enable their automatic detection in health records.

- Context-based identification: the data element is identified based on the given context, e.g., each string located next to the string 'Social Security Number' has a high potential of being a social security number. Together with the defined synonyms of personal and medical data elements we use state-of-the-art document analysis engines to automatically identify personal and medical data more reliable.

*Transformation:* Clinical research frequently utilizes proprietary data formats that are often incompatible with the data standards of other organizations. As a result, clinical data can rarely be exchanged between different organizations [42]. The purpose of the transformation phase is to convert the structured personal and health data into standardized data formats. Due to the complexity of standards such as HL7 or CDISC, we developed appropriate mapping schemes to ensure standard compliance of the generated output. Standard data formats, such as HL7 CDA, consist of a header and a body. The header includes the context in which the document was created, and the body contains the actual content of the document. The purpose of the header is to support communication across and within institutions, facilitate clinical document management, and facilitate the compilation of an individual patient's clinical documents into a lifetime electronic health record. MEDSEC guarantees that

- the body of a CDA document (either an unstructured blob or a structured markup) does not include any personal data, and that
- all information, needed for further processing of the data is included in the header without reducing privacy.

*Pseudonymization:* A server-side instance (e.g., HSM) acts as cryptographic module for executing the necessary cryptographic steps within a trusted secure environment. The cryptographic operations include all encryption and decryption operations required for functions, such as user authorization and authentication. The client-side cryptographic operations, required, e.g., for the challenge/response-style authentication procedure, are carried out with the user-owned security token doubling as secure keystore for the authentication credentials and a client-side cryptographic module. The architecture (see Figure 2) is realized as a multi-tier hull model with three different layers. Each layer is responsible for one step in the data access process. The user has to pass all layers in order to retrieve the actual health records. The outer hull, the authentication layer, is responsible for authenticating the user by requiring him to prove his identity. Technically, the outer hull is realized by the outer asymmetric keypair (outer public key OPuK and outer private key OPK) that is stored on the user's security token. The keys on the security token are only accessible when entering the correct PIN, thus providing two-factor authentication. Authentication involves the user's and the server's outer keypair, the user's internal user ID (IUID), and a random value. The user's outer private key is also used to decrypt his inner private key, which in turn is needed for decrypting the inner symmetric key. The inner symmetric key (ISK) and the inner private and public keys (IPK and IPuK) form the inner hull, the authorization layer. Without the inner symmetric key, the user cannot access the correct pseudonyms which are encrypted with his

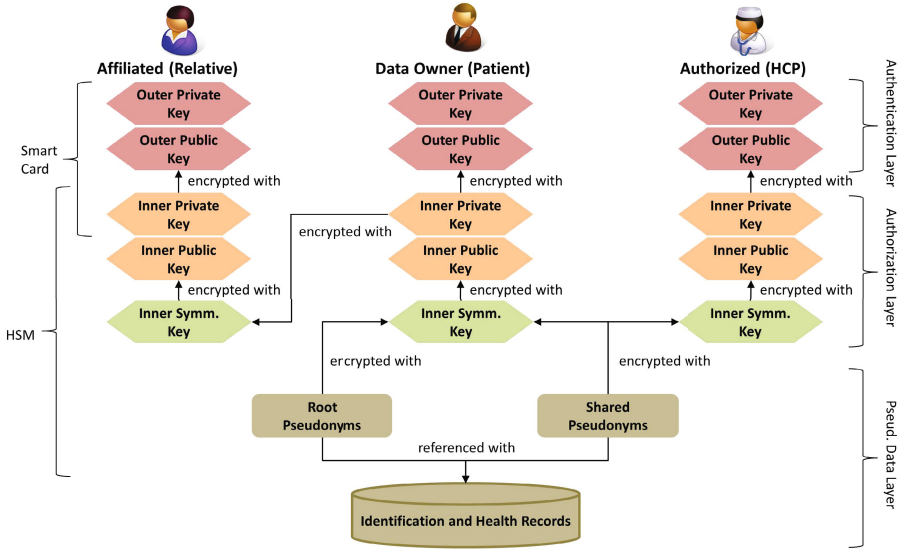


Fig. 2. Pseudonymization architecture ([36])

inner symmetric key. The pseudonyms could be directly encrypted with the inner public key and would still be secured against unauthorized access. However, defining an additional inner symmetric key has the following advantages: As symmetric encryptions are executed faster than the costly asymmetric cryptographic operations, reducing the number of encryptions/decryptions involving the asymmetric keys increases the overall execution speed. At the same time, it prevents the user from directly accessing the inner symmetric key, as it is only present in plaintext within the secure environment of the HSM where the pseudonyms are encrypted and decrypted. The plaintext pseudonyms are attached to the actual health records, and both together represent the innermost

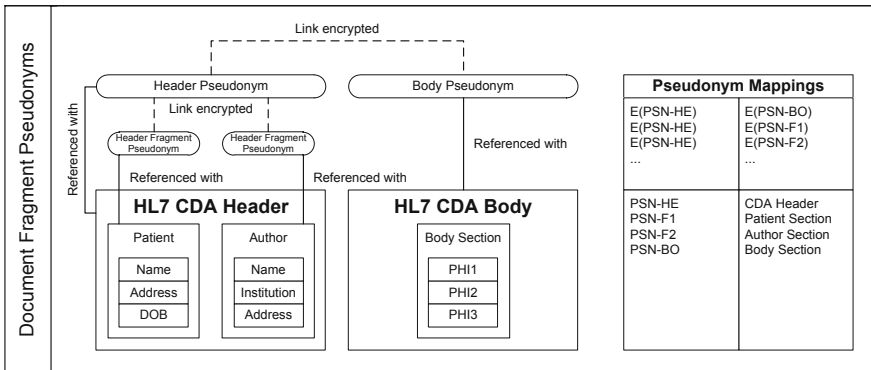


Fig. 3. Document Fragment Pseudonyms



layer, the concealed data layer. Figure 3 demonstrates the pseudonyms of HL7 CDA documents that are fragmented into several sections: While the CDA body section is assigned a single body pseudonym, the CDA header section is further fragmented (into header, patient, and author sections) and each fragment assigned an individual fragment pseudonym. As can be seen in the Pseudonym Mappings, the pseudonyms are attached to the CDA document fragments in plaintext, while the links between the fragments are effectively concealed by encryption. Thus, if in possession of the correct decryption key, the mappings can be decrypted and thus the links between the fragments restored.

## 4 Conclusion

MEDSEC was implemented into a software solution and tested within a national health-care provider in Austria that treats about 250.000 inpatients and 600.000 outpatients annually. Initial test runs with a limited document base demonstrated the system's practicality, producing promising results. The system is currently undergoing a test run on a larger scale with minor modifications to further improve the system, especially concerning the quality of the OCR and data identification output. The results will be presented in detail in a future publication. The project results enable to strengthen clinical research and harbor considerable economic benefits for the society due to the decreased treatment costs and more efficient clinical trials: MEDSEC simplifies the analysis of medical data by providing more representative samples and, thus, reduces the time required for carrying out clinical research (including clinical trials). This has two major advantages: (i) Clinical research can be carried out in a fraction of the (original) time due to faster recruitment. This is a powerful argument, because research organizations rely on the fast publication of research results. (ii) A larger sample results in more reliable and significant outcomes and has a major influence on the research quality. Digitized health records reduce costs for hospitals and research organizations in the following ways: (i) They save expensive archive space of paper-based health records. (ii) Digitization has the side effect of allowing the categorization of data and, thus, the fast and efficient search for specific information, which results in improved treatment processes for the patient. (iii) The conversion of medical data into standard formats, such as HL7, allows the more efficient administration and use of this data in clinical environments.

**Acknowledgments.** The research was funded by BRIDGE (#824884) and by COMET K1, FFG - Austrian Research Promotion Agency.

## References

1. Ernst, F.R., Grizzle, A.J.: Drug-related morbidity and mortality: Updating the cost-of-illness model. *Journal of the American Pharmacists Association* 41(2), 192–199 (2001)
2. Pope, J.: Implementing EHRs requires a shift in thinking. PHRs—the building blocks of EHRs—may be the quickest path to the fulfillment of disease management. *Health Management Technology* 27(6), 24 (2006)
3. Maerkle, S., Koechy, K., Tschirley, R., Lemke, H.U.: The PREPaRe system – Patient Oriented Access to the Personal Electronic Medical Record. In: *Proceedings of Computer Assisted Radiology and Surgery, Netherlands*, pp. 849–854 (2001)

4. Masi, J.D., Hansen, R., Grabowski, H.: The price of innovation: New estimates of drug development costs. *Journal of Health Economics* 22, 151–185 (2003)
5. 2000, C.I.: R&D Briefing: Benchmarking for Efficient Drug Development (2000)
6. Anton, A.I., Earp, J.B., Reese, A.: Analyzing website privacy requirements using a privacy goal taxonomy. In: *Proceedings of the IEEE Joint International Conference on Requirements Engineering*, pp. 23–31 (2002)
7. Squicciarini, A., Bertino, E., Ferrari, E., Ray, I.: Achieving privacy in trust negotiations with an ontology-based approach. *IEEE Transactions on Dependable and Secure Computing* 3(1), 13–30 (2006)
8. W3C: Platform for Privacy Preferences (P3P) Project (October 2007), <http://www.w3.org/P3P/>
9. Pfitzmann, A., Koehntopp, M.: Anonymity, Unlinkability, Unobservability, Pseudonymity, and Identity Management – A Consolidated Proposal for Terminology. LNCS. Springer, Heidelberg (2005)
10. Taipale, K.A.: Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy and the Lessons of King Ludd. *International Journal of Communications Law & Policy* 9 (2004)
11. Peterson, R.L.: Patent: Encryption system for allowing immediate universal access to medical records while maintaining complete patient control over privacy. US Patent US 2003/0074564 A1 (2003)
12. Thielscher, C., Gottfried, M., Umbreit, S., Boegner, F., Haack, J., Schroeders, N.: Patent: Data processing system for patient data. Int. Patent, WO 03/034294 A2 (2005)
13. de Moor, G.J., Claerhout, B., de Meyer, F.: Privacy enhancing technologies: the key to secure communication and management of clinical and genomic data. *Methods of Information in Medicine* 42, 148–153 (2003)
14. Gulcher, J.R., Kristjánsson, K., Gudbjartsson, H., Stefánsson, K.: Protection of privacy by third-party encryption in genetic research. *European Journal of Human Genetics* 8(10), 739–742 (2000)
15. Pommerening, K.: Medical Requirements for Data Protection. In: *Proceedings of IFIP Congress*, vol. 2, pp. 533–540 (1994)
16. Pommerening, K., Reng, M.: Secondary use of the Electronic Health Record via Pseudonymisation. In: *Medical and Care Compunetics* 1, pp. 441–446. IOS Press (2004)
17. Dolin, R.H., Alschuler, L., Beebe, C.: The hl7 clinical document architecture. *J. Am. Med. Inform. Assoc.* 8(6), 552–569 (2001)
18. Fischer-Huebner, S.: *IT-Security and Privacy: Design and Use of Privacy-Enhancing Security Mechanisms*. Springer (2001)
19. European Union: Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities* L 281, 31–50 (1995)
20. Hinde, S.: Privacy legislation: a comparison of the US and European approaches. *Computers and Security* 22(5), 378–387 (2003)
21. Hornung, G., Goetz, C.F.J., Goldschmidt, A.J.W.: Die künftige Telematik-Rahmenarchitektur im Gesundheitswesen. *Wirtschaftsinformatik* 47, 171–179 (2005)
22. U.S. Department of Health & Human Services Office for Civil Rights: Summary of the HIPAA Privacy Rule (2003)
23. U.S. Congress: Health Insurance Portability and Accountability Act of 1996. 104th Congress (1996)
24. Schabetsberger, T., Ammenwerth, E., Göbel, G., Lechleitner, G., Penz, R., Vogl, R., Wozak, F.: What are functional requirements of future shared electronic health records? *Connecting Medical Informatics and Bio-Informatics*, 1070–1075 (2005)

25. Riedl, B., Neubauer, T., Goluch, G., Boehm, O., Reinauer, G., Krumböck, A.: A secure architecture for the pseudonymization of medical data. In: Proceedings of the Second International Conference on Availability, Reliability and Security, pp. 318–324 (2007)
26. United States Department of Health & Human Service: HIPAA Administrative Simplification: Enforcement; Final Rule. Federal Register / Rules and Regulations 71(32) (2006)
27. Council of Europe: European Convention on Human Rights. Martinus Nijhoff Publishers (1987)
28. Maris, K.: The Human Factor. In: Proceedings of Hack.lu, Luxembourg (2005)
29. Thornburgh, T.: Social engineering: the “Dark Art”. In: Proceedings of the First Annual ACM Conference on Information Security Curriculum Development, pp. 133–135. ACM Press (2004)
30. Schmidt, V., Striebel, W., Prihoda, H., Becker, M., Lijzer, G.D.: Patent: Verfahren zum Beden oder Verarbeiten von Daten. German Patent, DE 199 25 910 A1 (2001)
31. Fraunhofer Institut: Spezifikation der Lösungsarchitektur zur Umsetzung der Anwendungen der elektronischen Gesundheitskarte (2005)
32. Caumanns, J.: Der Patient bleibt Herr seiner Daten. Informatik-Spektrum, 321–331 (2006)
33. Heurix, J., Karlinger, M., Neubauer, T.: Pseudonymization with metadata encryption for privacy-preserving searchable documents. In: Proceedings of the 45th Hawaii International Conference on System Sciences, HICSS 45 (2012)
34. Heurix, J., Karlinger, M., Schrefl, M., Neubauer, T.: A Hybrid Approach integrating Encryption and Pseudonymization for Protecting Electronic Health Records. In: Proceedings of the Eighth IASTED International Conference on Biomedical Engineering, p. 117 (2011)
35. Heurix, J., Neubauer, T.: Privacy-Preserving Storage and Access of Medical Data through Pseudonymization and Encryption. In: Furnell, S., Lambrinouidakis, C., Pernul, G. (eds.) TrustBus 2011. LNCS, vol. 6863, pp. 186–197. Springer, Heidelberg (2011)
36. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. International Journal of Medical Informatics 80(3), 190–204 (2011)
37. Neubauer, T., Kolb, M.: An Evaluation of Technologies for the Pseudonymization of Medical Data. In: Lee, R., Hu, G., Miao, H. (eds.) Computer and Information Science 2009. SCI, vol. 208, pp. 47–60. Springer, Heidelberg (2009)
38. Neubauer, T., Riedl, B.: Improving patients privacy with pseudonymization. In: Proceedings of the International Congress of the European Federation for Medical Informatics (2008)
39. Riedl, B., Grascher, V., Fenz, S., Neubauer, T.: Pseudonymization for improving the privacy in e-health applications. In: Proceedings of the Forty-First Hawai’i International Conference on System Sciences (2008)
40. Riedl, B., Grascher, V., Neubauer, T.: A secure e-health architecture based on the appliance of pseudonymization. Journal of Software (2008)
41. Hendry, M.: Smart Card Security and Applications, 2nd edn. Artech House, Inc., Norwood (2001)
42. Waegemann, C.: Status report 2002: Electronic health records. Medical Records Institute, Boston (2004)

# A Detailed Process Model for Large Scale Data Migration Projects

Klaus Haller<sup>1</sup>, Florian Matthes<sup>2</sup>, and Christopher Schulz<sup>2</sup>

<sup>1</sup> Swisscom IT Services Finance AG, Testing & QA, Pflanzschulstraße 7,  
8004 Zürich, Switzerland

<sup>2</sup> TU München, Lehrstuhl für Informatik 19 (sebis), Boltzmannstraße 3,  
85748 Garching b. München, Germany  
Klaus.Haller@swisscom.com,  
{Florian.Mattes, Christopher.Schulz}@in.tum.de

**Abstract.** Data migration projects might sound exotic. They are not. Instead, they are common in all medium and large enterprises, e.g., when replacing applications with new (standard) software or when consolidating the IT landscape in the aftermath of mergers and acquisitions activities. General-purpose methodologies such as Scrum focus on managing projects. However, they do not discuss (data-migration) domain-specific tasks. These tasks are the focus and contribution of this paper. It elaborates and compiles them into a process-model. The model defines the logical and temporal dependencies between the tasks and clarifies roles and responsibilities in a migration project. Thereby, the variety of used know-how-sources sets this paper apart from any previous work in this area. We synthesize not only existing literature and own project experience in the German automotive and the Swiss banking sector. We also incorporate the results of twenty-five qualitative interviews from various industry sectors guaranteeing a high validity and applicability of our findings.

**Keywords:** data migration, project management, process model, deliverables.

## 1 Introduction

Data migration is a software-supported one-time process migrating data from a (supposed to be shut down) source to a target application with a typically different data model. It might sound exotic for many software engineers, but market shares prove the opposite. Application and system software development account for 9.35% of the global software industry. In contrast, the IT services industries account for 90.65% [1]. It centers on how to run applications or to commission applications successfully on the customer site. One aspect is data migration<sup>1</sup>. Large scale data migration projects which move a high volume of data while involving many different stakeholders have two main triggers. The first trigger is application replacements.

---

<sup>1</sup> In general, data migration competencies are seldom needed, often resulting in missing internal resources [4]. Moreover, migration tasks add to the daily-work load of staff. Together with being a non-business enabling project, this makes it sensible to rely on service providers.

The data stored in the old system is transferred or migrated into the new one [2]. The second trigger is Mergers and Acquisitions (M&A) activities. In the process of consolidating redundancies, the data from the applications to be shut down has to be migrated to the “surviving” ones [3].

Comparing application development and data migration with sports means comparing a marathon with a 100 meters sprint race. Applications are long-living. They comprise thousands or millions lines of code coded by creative minds. Small mistakes are accepted and are corrected in later releases. In contrast, data migration comprises migration scripts, each possibly a few hundred lines of code. However, like a 100 meter sprint race, every small mistake leads to a complete failure. For instance, if a bank loses 0.1% of its customers during data migration, it is catastrophic. Thus, data migration is tested to be 100% correct or it is postponed. On the other side, nobody cares about deliverables after the data migration is completed. More technical details and challenges are discussed in previous publication [4, 5].

The aim of this paper is to take a project management perspective on data migration. Due to the needed high stability and reliability, we focus particularly on the process model and the various deliverables common to all projects. Complementing general-purpose project methodologies such as Scrum [6], we address two research questions:

Q1: Which process models for data migration do academia and industry suggest?

Q2: How does a detailed process model incorporating roles, deliverables, and phases for large scale data migration projects look like?

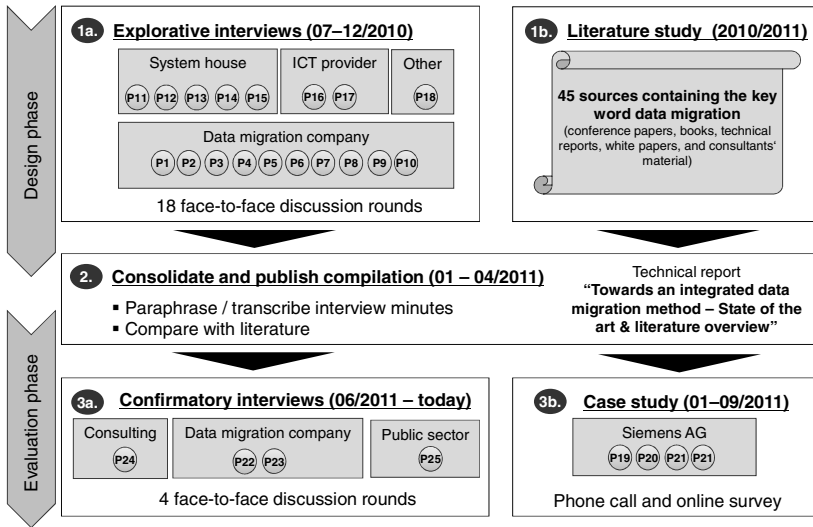
Throughout this paper we synthesize the various separate perspectives of the authors and their previous publications. The experiences come from the Swiss banking industry [2, 4], from data migration in the context of M&A projects [3], and from an interview-based study on the topic in various industries [7]. Additionally, this paper complements a previous joint paper on risk and quality assurance measures in data migration projects [5] by focusing now on processes for reducing quality risks.

The remainder of the paper is structured as follows: Section 2 addresses our research method. Section 3 outlines seven data migration approaches as found in academia and practice. We present the process model in Section 4 followed by an overview on the evaluation activities in Section 5. Finally, Section 6 concludes the paper with a short summary, a critical reflection, and an outlook.

## 2 Research Method

The goal of our research is the design of a detailed process model for data migration projects, i.e., an artifact which covers a broad scope while possessing a high level of detail. It must contain all important phases, roles, and deliverables. For understanding, executing, and evaluating the research we applied a design science approach adhering to the guidelines proposed by Hevner et al. [8]. The innovative artifact is designed to meet the business need, i.e., accessing semantically identical data with a different application. Figure 1 illustrates the artifact’s input for the design and evaluation phase. For confidentiality reasons, we have omitted the real names of the experts

(indicated via gray circles and a sequence number) and their organizations. An intermediary result was the technical report [7] marking the transitioning from design to evaluation. In the following, Hevner's guidelines are shortly discussed.



**Fig. 1.** Artifact's input for the design and evaluation phase

- Design as an artifact: Our process model is a method, thus a viable artifact being produced as part of the research. It guides data migration project participants in how to migrate data from a source to a target database.
- Problem relevance: A technology-based solution to migrate data is important as elaborated in the introduction section and literature (e.g., [2, 9]).
- Research contribution: To our knowledge, there is no detailed process model for data migration. The related work section contains a subset of sources we examined in 2010/11 by means of a systematic literature analysis. The study comprised 45 sources centering on data migration. It was conducted in parallel to the surveys and is partly published in [10]. Thus, the model provides a clear and verifiable contribution in the area of data migration.
- Research rigor: Besides the literature study, 18 data migration experts from six different companies in German speaking countries have been interviewed in half-day discussion rounds and on-site workshops between July and December 2010. The partners have been selected due to their domain knowledge following the idea of theoretical sampling<sup>2</sup>. The findings gained during these exploratory expert surveys have been systematically documented, transcribed, and consolidated. The result was a technical report [7] serving as a design foundation for our final process model.

<sup>2</sup> Theoretical sampling was introduced in the context of social research describing the process of choosing new research phenomena and to gain a deeper understanding [11]. It can be considered as a variation of data triangulation: using independent pieces of information to obtain better insights on something that is only partially known or understood.

- Design as a search process: Besides literature survey and interviews, the elements of the model date back to findings which we made in the course of several data migration projects in the financial industry. Achieved results have been previously published (cf. [2, 4, 5]).
- Communication of research: The process model is published in this paper to make it available to both management and technology.
- Design evaluation: Even though the model has its roots in industry experience we evaluated the artifact by means of four confirmatory interviews and a case study. Details are provided in Section 5 of this paper.

### 3 Related Work

Comprising more than 45 sources specifically centering on data migration, our literature study revealed that the topic is mainly addressed by practitioners. Most recent publications originate from consulting companies and tool vendors (e.g., [12, 13]). For this group, a data migration methodology is a competitive advantage. Hence, their approach often lacks details. This Section outlines the seven most detailed models from academia and industry focusing on their phases, deliverables, and roles.

Ground breaking work from academia is the PhD thesis of Aebi [9]. Subdivided in an explorative and execution part, the author presents the five-phase data migration model MIKADO. While the phases of problem analysis and definition are only conducted in the explorative step, preparation for and in the intermediate and target system also take place during execution. His work is technology-independent, proposes an intermediary staging area, and is supplemented by a tool architecture named DART. Describing the process of data migration on a general level Aebi refrains from pointing out concrete roles and deliverables.

Designed at Trinity College in 1997, the Butterfly Methodology is a method for migrating legacy information systems in a mission-critical environment [14]. Comprising six phases, the approach focuses on reducing the downtime of source and target application. The core idea is based on temporary databases which are sequentially set up, filled, and migrated to the target, before the next iteration starts (the number of data items declines over time). Presented phases outline key activities on a very high level only. The work centers on technical aspects, and, consequently, leaves out deliverables and roles of a data migration.

In their book Brodie and Stonebraker propose the Chicken Little strategy [15]. It is an incremental method which lets legacy and target systems interoperate during migration by using a gateway as a software mediating module. Pursuing a stepwise approach where the legacy applications are gradually rebuilt on the target platform, the strategy resolves into 11 briefly explained steps. Since emphasis is put on legacy migration in its entirety, only step 8 treats data migration details. Consequently, activities, deliverables, roles, or management aspects are not discussed. Finally, Wu et al. remark that gateways lead to the increase of the migration complexity [14].

John Morris looks at data migrations from a project management perspective based on various commercial data migration projects he was involved in [16]. The author elaborates on the stakeholders and key concepts of each data migration project. Bearing the name of his book, Morris' approach consists of four stages: project

initiation, data preparation (2x), and build, test, and go live. He defines precisely the activities and deliverables of each of those stages as well as the stakeholder in charge.

In cooperation with Informatica, The Data Warehouse Institute (TDWI) released a monograph about best practices in data migration [17]. Its main contribution consists in a seven-phase data migration process which covers several organizational and technical recommendations and is backed by a tool survey (Informatica, obviously, is a vendor in this area). Again, the author Philip Russom does not provide roles and deliverables needed in data migrations.

In 2007 Endava issued a white paper on key practices for data migration projects [12]. Targeted at migration project managers, it explains a set of key practices based on an 8-phase process. What sets this paper apart from others is the description of four roles: data migration project manager, data migration architect, business analysts, and testers. Endava provides a solid starting point without discussing technical, more data oriented aspects, specific tasks and roles, or deliverables.

Accenture advocates breaking down a migration project into a series of well-defined atomic level tasks, control metrics, and procedures that reduce cost and time to completion [13]. The distinctive feature of the work consists in the elaboration of concrete deliverables and roles (denoted key groups), even without giving a precise definition. In a nutshell, their process model has six phases which can be carried out in parallel. As a typical industry publication, the document remains on a very high level such that it only presents a short rundown on each phase. No refinements are made regarding roles and deliverables used during data analysis and transformation.

The above approaches represent a solid basis for the design of a detailed process model. However, our paper intends to go one step further. It provides empirically obtained information on phases, roles, and deliverables particular for data migrations.

## 4 A Process Model for Data Migration

In this section we describe the process model we devised rooted in on our own experience, statements from the data migration experts, as well as the literature study. The latter source is made explicit via references. Due to confidentiality reasons we cannot provide pointers to the industry experts. The success of a data migration project relies on many factors. One is clearly defined roles and teams. Below we list the most important ones, their tasks and responsibilities.

- The **business/program sponsor** initiates the migration project [13]. She/he clarifies the project scope, represents the business needs, and ensures the funding. Furthermore, the sponsor supervises the teams' project manager.
- The **customer core team** takes up the client's perspective. Many team members are business domain experts. They know the source application and understand how it contributes to the execution of the day-to-day business [16]. For this reason, they should be also employed as testers.
- The expertise of the **target application team** is the target application [15]. If the latter is replaced by a new one, building up a knowledgeable target application team might become a challenge. If there is already a target, the application management team and key users have to be involved [2].



- **Data migration team** is responsible for analyzing the data migration needs, implementing transformation rules, as well as running and testing them [12, 16]. This paper mainly focuses on their tasks.
- The **infrastructure and logistics team** provides infrastructure management services for the overall project [4]. They take over tasks such as managing the network, servers, database, or access rights. Certainly, the data migration team might manage specific data migration tools by themselves
- Some projects require the participation of **external auditors** for regulatory reasons [16]. In other cases, they participate at the discretion of the company. Auditors assess the process, project setup, migrated data, and reports. Their focus might not be on the migration alone but on the replacement project.

The process model (Figure 2) consists of four stages which are divided into phases yielding different deliverables. Stage one is the initialization phase for setting up the project organization and the technical infrastructure. Stage two is the development and stage three the testing phase. Often, both require several iterations until the migration is mature. This characteristic is accounted for with the help of the white arcs. In stage four the target application is put into operation.

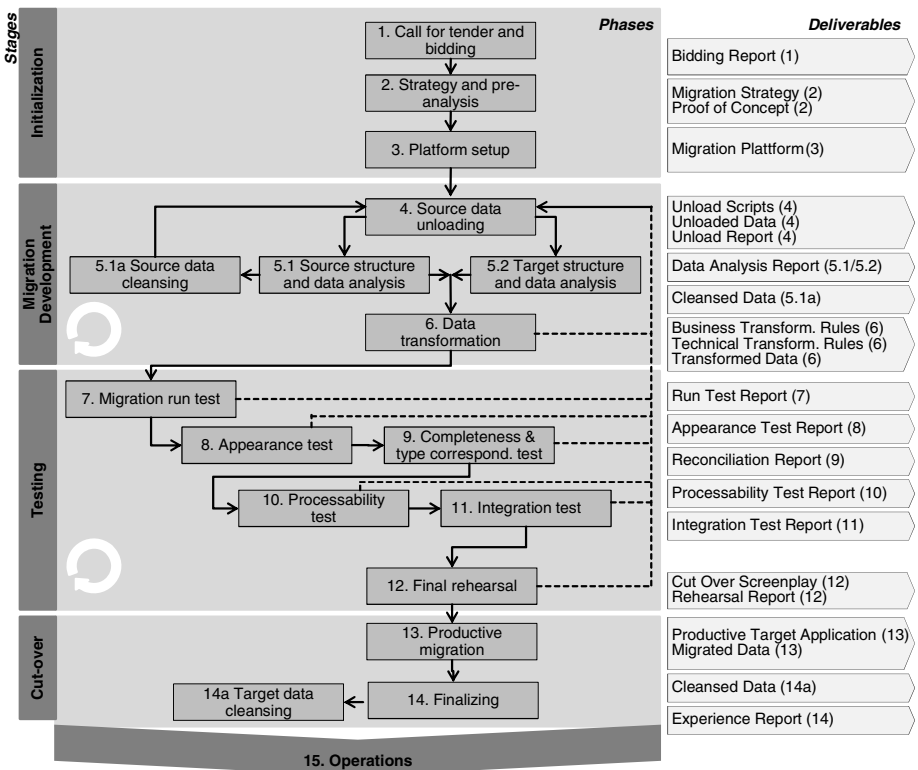


Fig. 2. Data migration process model

## 4.1 Stage 1: Initialization

**Call for Tender / Bidding (1).** Initially, it is not clear whether the newly formed IT department or an external IT service provider (and if so, which one) will take care of the data migration. This is the outcome of the first phase. Based on a customer questionnaire, preliminary data analysis, and their experience, the external company elaborates a bidding report. Pointed out as a necessity by our interview partners, the document clarifies the required resources, various risks, initial project planning, and the costs [7]. Most importantly, it covers the decision as to who has to take part in the project. Senior experts and managers from the company as well as from external service providers are involved in this phase.

**Strategy and Pre-analysis (2).** The primary outcome of this phase is a migration strategy defined by the data migration team together with the customer core and target application team. It refines the initial plan of the bidding report (e.g., big bang versus incremental approach [15] or how to cope with historic data in the source application [12]). The migration strategy covers also the organizational and technical constraints such as large data volumes, a target database already in use, proprietary interfaces, poor data quality, or distributed project teams. Furthermore, it details measurable acceptance criteria defining when the project is actually done. Depending on the time and the budget, a proof-of-concept with a small data set helps to raise the project morale and management support while allowing for a better cost estimation. A first prototype was mentioned by many of our interview partners [7].

**Platform Setup (3).** The data migration and infrastructure teams jointly set up the data migration platform and infrastructure respectively. While the former group is in charge of migration specific tools, the latter takes care of hardware, operating system, network, databases, etc. The data migration platform pursues two goals. First, it serves the data migration team for developing and testing the data migration scripts. Second, the test and the final migration run on this platform. The platform encompasses a staging area, an execution framework for orchestrating large numbers of data migration programs (semi-)automatically, a repository for the migration programs, and tools for developing data migration programs (cf. [2, 9, 17]).

## 4.2 Stage 2: Migration Development

**Source Data Unloading (4).** Unloading data from source to the migration platform must not hinder the daily operations on the still-in-use source application. Also, the unloaded data must be consistent over all modules. Often overlooked in the literature but valued as important among our interview partners, the unloading phase requires careful planning as to when and how to extract the data. After having planned the unloading, the customer core team (respectively the application management team) maps source and target data structures to the staging area, implements scripts (unloads are needed several times), and finally initiates the extract. The copying process might include some high level filtering, e.g., copy only master data and non-transactional data (cf. [2]). Furthermore, additional data not stored in the source but required by the target might be added, i.e., enriching. The outcome of this phase are the unload scripts, the actual (enriched) data downloaded to the migration platform unload

staging area, and a report with the number of unloaded data items. The latter is important to make sure that all data is going to be moved to the target.

**Source/Target Data and Structure Analysis (5.1/5.2).** The data models of the source and target application differ. Transformation rules implemented in the data migration scripts overcome this gap, but a profound understanding of the source and target data and structure is indispensable in beforehand. The analysis is mainly implemented by the data migration team [12]. Regarding data semantics and application knowledge, the team relies on the input of the customer core and target application team. In close cooperation, the three may unveil data quality issues which they document and assess. The outcome is a data analysis report which also includes necessary repairing work. As outlined by our interviewees, the document describes the source/target structure and the quality of the in-scope data on an attribute level [7].

**Source Data Cleansing (5.1a).** Any previously detected data quality issues should be cleansed to reduce the risk of future errors [16]. This can take place in the source application, during the transformation, or in the target application [12]. The latter two options entail the risk of more complex transformation rules. Therefore, cleansing should be performed directly in the source conjointly with the customer core team. In this way, the cleansing can be handled separately and performed by users of/on the source application. Final result is cleansed data meeting the target's requirements.

**Data Transformation (6).** The data analysis report compiled during phase 4 enables specifying and implementing the transformation rules. The interview partners explained that this is a two-step process [7]. First, the customer core team and the data migration team jointly formulate business transformation rules. These are an informal specification of the mapping of business objects types (e.g., customer, account) as well as tables and attributes between the source and target data structure. Second, the data migration team describes, implements, and tests technical transformation rules, e.g., using SQL or PL/SQL for relational databases. These rules also include data harmonization logic accounting for the data already residing in the target.

### 4.3 Stage 3: Testing

**Migration Run Test (7).** Before one can test migration scripts in detail, the infrastructure must work and the ordering of the migration of the business objects types must be defined. Furthermore, during the final productive migration, all data programs have to run smoothly within an acceptable time frame and respecting dependencies. Consequently, a migration run test, performed by the data migration team, validates exactly this. There are full and partial migration run tests. Full run tests validate the migration of all data [12]. If the infrastructure remains identical, the execution time should be the same as for the final migration. Migrating only parts of the data speeds up the migration but comes with higher consistency risks [5]. The deliverables is a run test report. It contains the logs with all error messages, crashes, and – in case of a full migration run test – the required execution time.

**Appearance Test (8).** The data in the source and target application must be semantically equal. The challenge is that both the data model and the representation of the data on the Graphical User Interface (GUI) differ between the source and target

application [2]. Business domain experts of the customer core team define a diverse set of data items of various business object types. On GUI level, they compare the test data items after each test migration (cf. [5] for a discussion on the risks of automating appearance tests). The outcome is an appearance test report. From a GUI perspective, it states for each data item of the test set whether it was migrated correctly.

**Completeness and Type Correspondent Test (9).** Data migration projects handle large amounts of data. The volume demands automation of the comparison of the data in the source and target, i.e., a reconciliation [4, 12]. Reconciliation scripts map primary keys (respectively identifiers) of the source and target applications. Thus, they identify whether data items got lost or emerged during the transformation and migration. The reconciliation is the only test covering all data items. This makes it sensible to consider also the types of objects on a high level. The implementation of the reconciliation and its execution after each migration test run is the task of the data migration team. Final deliverable consists in a reconciliation report containing all data items lost during the migration as well as all data items being emerged from nowhere.

**Processability Test (10).** Completeness and appearance on GUI level do not guarantee that the data can be processed faultlessly. There might be mandatory items which are missing and for which the GUI presents default values instead. However, the business logic of the target application might crash when processing the incomplete and/or default data. Other causes for problems are target application parameterizations being not compatible to the migrated data. The processability test addresses these challenges. The customer core and the data migration team specify which processes have to be tested and what kind of test data is leveraged for that. The tests can be executed manually on GUI level. In addition, application specific batch processes have to be applied (e.g., end-of-day or end-of-year processing). The result is a processability test set with batch processes and data items for their execution and a processability test report stating which processes could be executed correctly.

**Integration Test (11).** In most cases, the target application is embedded in an IT landscape consisting of a large number of interconnected applications. If the decision has been in favor of the migration, the business processes supported by the target application (including migrated data) and all connected applications must still function. This requires an integration test phase checking whether the references between the data in the respective target application and its neighbors work in both directions. Integration tests are end-to-end since they comprise the running of processes spanning over various applications including the newly migrated target application. The deliverables of the integration test phase are, first, an integration test set. It consists of a list of processes to be executed including the data items to be used. While the customer core and the data migration team have to work out this list together, only the former is required to carry out the tests. The second outcome is an integration test report stating which of the tests succeeded and failed respectively.

**Final Rehearsal (12).** An unsuccessful migration is expensive and a risk for the reputation of the enterprise. Thus, several final rehearsals are a last testing measure [13, 16]. It runs under the boundary conditions (e.g., hardware, course of action, amount of data) of the productive migration while making use of the same transformation rules. More precisely, the final rehearsal executes the data migration

with the transformation rules (phase 6) and an unload from the source system (phase 4) [12]. Next, it covers all tests from phases 7-11 in a condensed form. Thereby, every member of the data migration and customer core team performs exactly the tasks as they do later during the cut-over, i.e., the productive migration. The deliverables are a finalized cut-over screenplay serving as a blue-print. After the rehearsal, the data migration team works out a rehearsal report with all test results and unexpected events. This allows the management to decide whether the productive migration can be initiated as planned or whether additional adjustments have to be made.

#### 4.4 Stage 4: Cut-Over

**Productive Migration (13).** The productive migration is the moment when the migration from the source to target application takes place [13, 15, 17]. Afterwards, the target application is released into production and the source is shut down. This is the point-of-no-return. A fall-back to the source application would be at expensive if impossible at all. Our interviewees referred to the convention of a final approval meeting with all team members [7]. The meeting decides whether the migration can start. The actual course of action consists of the tasks described in the phase 4, 6, and 7 to 11 and results in a productive and tested target which contains the migrated data.

**Finalizing (14).** The finalizing phase starts when the target application is up and running, i.e., all data has been successfully migrated and the end users work on the target [13, 16]. The target application team ensures adequate performance and stability. If this goal is achieved, the responsibility of the data migration team and all other project teams described above ends. They hand over the responsibilities to the application management team [17]. Our interview partners highlighted the importance of a short experience report compiled by the data migration team. It contains lessons learned which aim to ease future data migration projects [7].

**Target Data Cleansing (14a).** There are situations when not all data cleansing can be carried out in the source application or during the transformation. This is particularly the case for projects with a large amount of data and a very short time frame. Then, data cleansing should be done as soon as the target application is in production. This also comprises the manual adding of data, which could not be migrated automatically or was not originally stored in the source application.

## 5 Evaluation

As discussed in Section 2, the design of presented process model rests on three pillars. The first one is the authors' own experience and know-how gained during several data migration projects. The second is a literature review of 45 industry and research sources. Finally, the third pillar consists of interviews and workshops with 18 data migration experts already applying elements of the model.

Though the results being delineated in this paper largely originate from industry experience, we apply a two-step validation procedure. Firstly, we presented the model to experts who did not contribute to the original process design but are also well versed in data migrations. Since June 2011, we have organized dedicated meetings

involving four industry experts. Held at TU München or the expert's site, the meetings aimed at presenting, discussing, and validating the model. We captured the feedback via a confirmatory three-page questionnaire and incorporated most of the suggestions in the model. For instance, the partners convinced us to enhance our approach by focusing on concrete artifacts yielded in the course of a migration project. Also, they confirmed the model's suitability for large scale projects and proposed the inclusion of milestones, a fact that is accounted for by the four stages.

The second validation step consisted of an evaluation over a longer time period. In January 2011, we were introducing the model to Siemens AG being in charge of replacing the product life-cycle management (PLM) application of a German car manufacturer [10]. In detail, the application named Process Designer (PD) has to be replaced by TeamCenter Manufacturing (TCM) until 2016. Thereby, Siemens develops (and sells) both applications. TCM offers enhanced user management functionalities and allows for the versioning of its data. At the time Siemens approached us, the project was in the "Strategy and pre-analysis" phase seeking for an appropriate process model ensuring a smooth transition to the successor. Siemens asked for a sound methodology in addition to a literature survey of common data migration approaches. There was no data stored in the well-documented target. Although the documentation about the source's data is in poor condition, preliminary analysis revealed that the data itself is of very high quality. Carried out by a project team located at three sites, the migration will be conducted incrementally on a vehicle-basis. We outlined our process model to four Siemens experts during a 1.5h phone call. Afterwards, the employees had to fill out an online survey containing 50 confirmatory questions. The survey confirmed that our data migration process model is generally in line with the experience and expectations of the Siemens experts. Roles, deliverables, as well as the phases have been confirmed. The experts added that the cleansing could be integrated in the transformation activities.

## 6 Conclusion

In this paper, we presented a detailed process model for large scale data migration projects consisting of four stages. The model targets the needs of the IT service industry and its respective IT consultants. As external specialists, these consultants work on getting applications to run on their customers' IT infrastructure. The task covers the migration of data into the new applications. Listed deliverables defined for each phase provide guidance in this highly risk-averse domain.

This contribution centered on the fundamental challenges of data migration projects. Many more questions remain open and have to be addressed during future research. In our opinion, the most pressing ones are:

- Coping with migration series, e.g., when replacing several systems. Business needs the freedom to determine the exact date and portion of data being migrated while working on a consistent application landscape.
- Analyzing the cost-drivers of data-migration projects. This also encompasses questions like which phases and deliverables are especially important and which organizational choices help keeping overall costs low.
- Evaluating the model for various storage paradigms. Certainly, the process model is designed independently of the underlying database technology.

However, we have to confess that our sources and experiences mainly originate from projects with (object-)relational databases. Thus, it would be interesting to see how the needs for projects with, among others, XML databases, object-oriented databases, or video and audio databases differ.

Existing questions should not hide the contribution of our paper: a detailed process model for data migration projects. It combines an academic approach with practical experience by synthesizing knowledge sources ranging from own experience, literature, and migration experts. Finally, the artifact endured an intensive evaluation.

**Acknowledgement.** We want to thank all industry partners who have contributed to the design and evaluation of the data migration process model.

## References

1. Segmental contributions of Global Software industry, <http://www.checkonomics.com> (last retrieved on January 9, 2012), The data reflects the market shares in 2005
2. Haller, K.: Datenmigration bei Standardsoftware-Einführungsprojekten. *Datenbank-Spektrum* 8(25) (2008)
3. Freitag, A., Matthes, F., Schulz, C.: M&A driven IT transformation – Empirical findings from a series of expert interviews in the German banking industry. In: *WI 2011, Zürich/Switzerland* (2011)
4. Haller, K.: Towards the Industrialization of Data Migration: Concepts and Patterns for Standard Software Implementation Projects. In: van Eck, P., Gordijn, J., Wieringa, R. (eds.) *CAiSE 2009. LNCS*, vol. 5565, pp. 63–78. Springer, Heidelberg (2009)
5. Haller, K., Matthes, F., Schulz, C.: Testing & Quality Assurance in Data Migration Projects. In: *ICSM 2011, Williamsburg/USA* (2011)
6. Schwabe, K.: SCRUM Development Process. In: *OOPLSA 1995, Austin, TX, October 15-19* (1995)
7. Matthes, F., Schulz, C.: Towards an integrated data migration process model - State of the art & literature overview. Technical Report. Garching b. München/Germany (2011)
8. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. *MIS Quarterly* 28(1) (2004)
9. Aebi, D.: Re-Engineering und Migration betrieblicher Nutzdaten. ETH Zürich (1996)
10. Donauer, S.: Ableitung & prototypische Anwendung eines Vorgehensmodells für die Migration von Product-Lifecycle-Daten bei einem Automobilhersteller. Diplomarbeit, TU München, Garching/Germany (2011)
11. Glaser, B., Strauss, A.: *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine, Chicago/USA (1967)
12. Endava: *Data Migration - The Endava Approach*, London/United Kingdom (2007)
13. Mohanty, S.: *Data Migration Strategies 1 & 2*. Information Mgmt. Special Reports (2004)
14. Wu, B., Lawless, D., Bisbal, J., Richardson, R., Grimson, J., Wade, V., O'Sullivan, D.: The Butterfly Methodology: A Gateway-free Approach for Migrating Legacy. In: *ICECCS 1997, Como/Italy* (1997)
15. Brodie, M.L., Stonebraker, M.: *Migrating Legacy Systems*, 1st edn. Morgan Kaufmann Publishers Inc., San Francisco (1995)
16. Morris, J.: *Practical Data Migration*, 3rd edn. British Informatics Society Ltd., Swindon (2006)
17. Russom, P.: *Best Practices in Data Migration*, Renton/USA (2006)

# Towards Automated *Generic* Electronic Flight Log Book Transfer

Carsten Kleiner and Arne Koschel

University of Applied Sciences & Arts Hannover  
Faculty IV, Department of Computer Science  
30459 Hannover, Germany  
{ckleiner, akoschel}@acm.org

**Abstract.** The automated transfer of flight logbook information from aircrafts into aircraft maintenance systems leads to reduced ground and maintenance time and is thus desirable from an economical point of view. Until recently, flight logbooks have not been managed electronically in aircrafts or at least data transfer from aircraft to ground maintenance system has been executed manually, since only latest aircraft types (e.g., Airbus A380) support electronic logbooks. This paper contributes concept and top level distributed system architecture of a *generic* system for automated flight logbook data transfer. It details a generic mapping component that facilitates flexible mappings between aircraft logbook systems as input and aircraft maintenance systems in the backend. Due to its flexible design the mapping component could also be used for different domains with similar requirements.

**Keywords:** system integration, data mapping, XML, aerospace engineering, generic interface, configurable mapping.

## 1 Introduction

Ground and maintenance time is very costly for airline operators, thus they try to minimize them for economical reasons. Today's still mostly manual transfer of flight logbook data from an aircraft into the operator's maintenance systems should be automated to get closer to achieving this goal. This will eventually reduce information transfer time and is likely to be less error prone as well. Thus in total it should result in reduced ground maintenance time and cost.

A generic automated flight log data transfer system needs to support the differently structured flight log information from different aircraft types and manufacturers on one side. On the other side different aircraft maintenance systems used by different operators have to be supported. Technically, all these systems are very likely distributed, even though typically in a common network within a single organization. Moreover, fault tolerance and (transactional) persistence to prevent data loss are required. Performance demands however, are not very critical due to a limited amount of log data per flight in practical applications.

To support those requirements, a *generic* transfer system for flight logbooks into maintenance systems needs to be designed and implemented. In a joint industry and



research cooperation (*Verbundprojekt*) the eLog system has been designed and prototypically implemented by the University of Applied Sciences Hannover in cooperation with Lufthansa Technik AG and edatasystems GmbH. Technically this system supports different – currently XML-based, but with heterogeneous XML schemata – versions of different aircraft flight log systems on the *input* side. On the *output* side different airline systems are supported, which may have different data models. As an example a relational DBMS with tables is used, that map (almost) 1:1 to the data structures of the *output* system. The *output* system is Lufthansa’s core aircraft maintenance system. Technically data is provided to the output system in an XML format closely reflecting the relational structure.

The resulting eLog system offers a generic distributed system architecture for the integration and mapping of the mentioned different XML-based flight log input data formats to different output aircraft maintenance systems. The mapping itself is configurable and quite flexible. Its only restriction is that input data has to be provided in XML files according to a predefined schema. However, since arbitrary entities can be mapped this does not affect the genericity of our approach at all.

Mapping of input objects to output objects is specified by an XML mapping document (conforming to a specific mapping schema) which is dynamically loaded into the mapping component. Information in the mapping document is also handled dynamically making the component extremely agile once in operation. In total all these features contribute towards our goal of a generic flight logbook tool. While in [13] a high level overview on eLog has been given, this paper’s main contribution are details on mapping options, their potential complexity and their implementation. These topics are discussed in section 4.

Initial tests of the prototypical implementation already validated the practical usefulness of the concept. Very few logbook data transfer systems exist. Those that do are flight operator internal and/or aircraft type and maintenance system specific.

Although concepts and approaches for application/data integration in general of course do exist – important ones are briefly discussed in section 2 – their application to flight logbook data is a novelty. To the best of our knowledge a *generic* flight logbook data transfer system has not been implemented before and is thus the key contribution of our overall work.

The remainder of this article discusses some related work, gives a high level eLog system overview, discusses as the main contribution the eLog mapping approach, and ends with a conclusion and some outlook to future work.

## 2 Related Work

Related work originates from different areas. *Conceptually* different enterprise application integration approaches [2,4,12] provide a potential foundation for the technical system architecture. The most important ones include: messaging based enterprise integration patterns [7], transactional DBMS based approaches [5,3], using an ESB/SOA as foundation [11], and finally an ETL-like data warehouse concept [8].

All of them potentially deliver feasible architectures for a system like eLog and have been evaluated (not shown here due to space constraints). Eventually a transactional

DBMS based architecture was chosen similar to an ETL approach. We omit a detailed discussion about the architecture selection at this point. The selection has been undertaken and is described in project internal documentation. It might be published by us in a future article. In this paper we will rather focus on the achieved results.

From an *application* point of view many systems exist, which transfer and map data from multiple input sources to different output sinks. E.g., Deutsche Post uses an Enterprise Service Bus for XML-based data transfer within a SOA [6].

A generic XML mapping architecture is discussed in [9]. Graph based mapping to transform structured documents is explored in [15]. SMART: A tool for semantic-driven creation of complex XML mappings is presented in [16]. Moreover (semi-)automated mapping of XML schemas has been discussed in many research papers, a pretty comprehensive survey is given in [14]. However, for the limited complexity of the XML schemas used in flight logbooks the project has decided that the overhead of employing a complex automated mapper, let alone choosing the most suitable one, is too big. Nevertheless the output of an automated mapper could be used as a base to define the mapping documents (cf. section 4.2) if our approach is applied to more complex domain schemas.

Looking at recent aircraft models in particular shows that only the latest models support electronic log books. In addition standardization of the data format is still in its early stages. Recently initial standardization has been designed in the so-called ATA specification [1]. This is quite helpful for the flight log input data within our work – although the ATA specification is still significantly in flux. One e-logbook tool for the aerospace industry is presented in [10]. But no *generic* flight logbook data transfer system has been documented yet.

### 3 eLog: System Overview

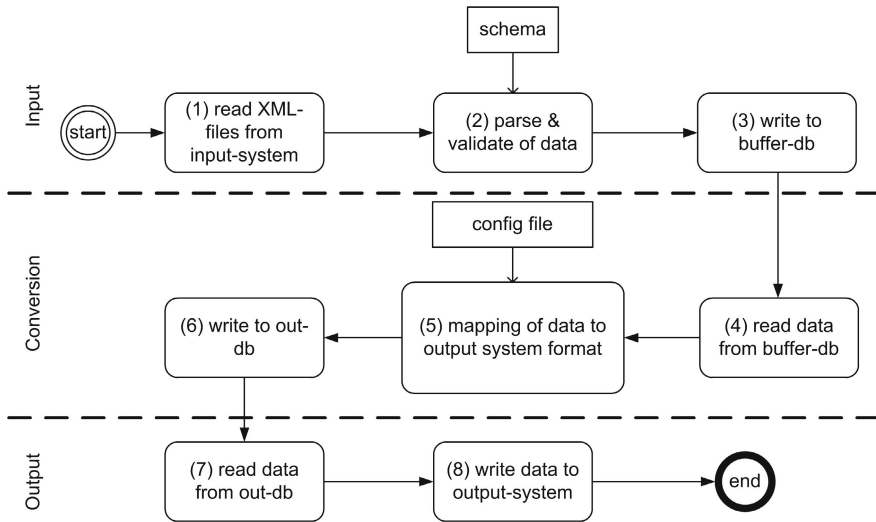
The designed generic flight log data transfer system is based on ideas frequently found in extract transform load (ETL) processes in data warehouse systems. Note though that the implementation itself is not based on data warehouses but rather uses a transactional DBMS-based approach as stated in the previous section.

In order to provide a brief eLog system overview we explain the data flow within eLog in the sequel.

#### 3.1 Data Flow within eLog

Data flow within eLog follows a sequence of steps to map the XML input data to arbitrary aircraft maintenance systems. Figure 1 shows the high level input data flow from the flight logbook system until procured to the maintenance system.

An input reader component – where different implementations for different source data formats may exist – uses polling (step 1) to check whether new XML files have been delivered by aircrafts. Polling is implemented by frequently checking a predefined directory on a dedicated server for newly added files. The data is validated against the XML schema (step 2) of the particular aircraft's electronic logbook format, e.g., against those defined in [1]. If the data is formally valid, it is transferred into a buffer database



**Fig. 1.** Generic eLog data flow

(step 3), where it is stored in its original format in an *XML-type* attribute. Else an error handling takes place. As long as the covered aircrafts provide source data in XML format a single database schema is sufficient for the buffer database.

Again using polling a mapping component checks for newly arrived source data (step 4) in the buffer database. It utilizes a flexibly configurable sequence of conversion functions to map the input data to a specific output target system (step 5); this step is explained in detail in section 4. The output data is stored in a database again (step 6). It closely resembles the data structure of the airline's maintenance system.

Consequently for each different maintenance system there will be an individual schema in the output database. Options in the mapping configuration include checks for dependent source information, flexible mapping as well as features to update existing entities in the output system; for details see section 4.

Eventually another upload component transfers data from the output database (step 7) into the airline maintenance system (step 8) whenever an entity of the maintenance system has been completely assembled. As one option the airline maintenance system used provided a Web services interface for programmatic access.

## 4 Mapping Rules

In this section the mapping of domain specific information between input and output systems is described in more detail. The different types of possible mappings will be presented first followed by a detailed description on how the mappings may be implemented. The implementation part will focus on the types of mappings that actually

occurred in our case study while the initial presentation is more general. Note though that the implementation itself is completely generic both in relation to domain or specific data format. That is because source and target entities, attributes and transformations are configured in the XML mapping description and may be from any arbitrary XML input or relational output. We have e. g. used these configurable mappings to deal with different system versions of input data; this will be explained in more detail in a future paper.

#### 4.1 Mapping of Information

In general the domains on the input side as well as on the output side will be structured according to some (already existing) data model. Typically the underlying data models can be modeled by the traditional ER approach. Matching general ER models is way too complex and also not necessary in most cases. In the situations considered in this paper input data will be available in XML structured files, i. e. mapping can be constrained to (potentially hierarchically structured) elements and their attributes on the input side. This assumption significantly reduces the complexity of the mapping process. In addition, the output side often requires XML structured information as well; this is even true for the case of relational database systems (RDBS) as in our case. RDBS trivially define a corresponding XML structure – tables correspond to XML elements with a table row being an instance of this element and the attributes for the different columns can be treated as XML attributes.

In conclusion and with the previously described assumptions we only have to deal with three levels of entities to be mapped: elements, attributes and attribute values. The details on these types of mappings will be described in the sequel with decreasing potential complexity.

**Mapping of Entities.** Mapping of entities may be further separated into different cases based on the number of entities on input and output side required for a single mapping operation:

- 1:1 mapping: the most common case of mappings when the data models are not too far apart. One XML element in the input file will be used to generate one XML element in the output. Details on the mapping will be defined on attribute (value) level and are explained below.
- 1:0..1 mapping: a special case of the previous mapping that has to be treated differently by the mapping process. One element in the input may lead to 0 or 1 element on the output side based on a mapping condition. The mapping process has to observe and understand the mapping condition and has to evaluate it. The condition evaluation may not be possible at the time the input element is provided, thus buffering and frequent re-checking of the condition is required in this case. If the element has to be mapped the details are determined as above.
- 1:n mapping: a single element on the input side generates multiple elements on the output side. This is also a very common case in the case studies that have been performed in our research. Such mappings are necessary due to different mapping

granularities between input and output data model. Details of such mappings are defined on the attribute level as before but in addition target entities have to be specified. Optional mapping of elements as in the 1:0..1 case are also possible as part of this mapping.

- m:1/m:n mappings: these mappings require multiple input entities to be present in order to be able to generate one (or more) entities on the output side. Similar to the 1:0..1 case these mappings require persistent buffering of input entities until all necessary information to generate the output entity are present. Alternatively one could perform the mapping directly and buffer the output entity until it is complete. We nevertheless decided to buffer the input entities. This proved handy in cases where a single input entity had to be used in several mapping steps.

**Mapping of Attributes.** Similar to the mapping of entities which – as described above – is typically defined in detail by a mapping of attributes there are different cases to be considered for mapping of attributes.

- 1:1 mapping: this is the most frequent case in our case studies where one attribute of the input system is mapped to one attribute in the output model. This kind of mapping may in addition use a simple transformation of the attribute values (see below) to take care of different representations of values in the two systems (e.g. date or address values).
- 1:n mapping: a single attribute is split into several attributes in the output model. In addition to the names of the output attributes the conditions on which to perform the split have to be specified here. Transformation of values is additionally possible for all of the output attributes.
- n:1 mapping: several input attributes are combined into a single output attribute. This kind of mapping did not occur in our case studies but should not pose a problem if the combination rules can be expressed with the mapping specification used. As all input attributes are present when this mapping is performed a simple application of the combination rules is sufficient.
- n:m mapping: this case also did not occur in our case studies. Similar to the case before this can also be treated by sequentially applying the individual mapping rules as long as the combination rules can be expressed. Explained in section 4.2 our mapping specification will define the mapping rules to be applied for each input entity and attribute. Thus the m:n case will lead to several mapping rule entries for each input attribute that can be applied one after the other.

**Mapping of Attribute Values.** In addition to mapping of entities and attributes we have already described that mappings on attribute value level are also possible. These may be performed by so-called converters, which are mainly used to accommodate different representations of the semantically same information between the domains. Such converters are typically simple to implement but occur in large numbers and thus are also a very important part of the mapping process. Details and examples of such converters can be found in section 4.2.

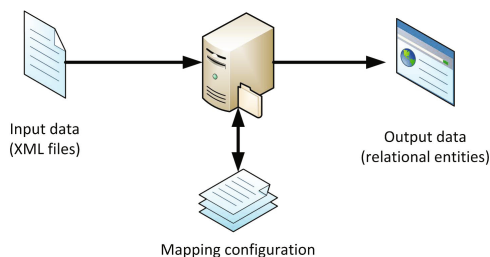


Fig. 2. XML based mapping process

## 4.2 Definition of Mappings

The general mapping process (cf. figure 2) is based on XML. Input information is assumed to be provided in XML files. In the case studies we received a single XML file per input entity, our system can also deal with multiple entities per file though. For each input entity there exists a mapping configuration document for the mapping process. This document, which is also an XML file following a certain schema, defines all necessary mappings for this input entity type. More details in this mapping specification will be explained in the following subsection.

**Specification.** The most interesting part of the mapping process is the specification of the mappings in XML files. As stated above one mapping document has to be generated per input entity type. Within this mapping configuration all possible mapping options presented in section 4.1 have to be taken care of. Whenever an input entity of the defined type arrives, the mapping configuration is analyzed and executed on this entity. Note that due to potential dependencies of certain parts of the mapping process on other input entities, which might not be present yet, a delayed execution of the mapping might be required. As the triggering event (arrival of the dependent entity) may occur at any time a frequent rechecking if the dependent entity is present has to be implemented. If immediate processing is required (which is not the case in our study) a notification mechanism would have to be employed instead.

A fragment of an example mapping specification is shown in figure 3. All output entities and their respective types that might obtain information from this input entity are specified and named for future reference. Apart from new entities (*init*) existing entities may be re-used and referenced (*update*) based on their ID information. In addition dependent entities that have to be present in order for the mapping to be executed may be specified and named. Currently only equi-joins based on attribute values of these entities are supported but more complex join types could be easily added if required.

Details of the mapping process as explained before are specified by attribute mappings within the entities. Consequently a mapping specification document as shown in figure 3 in addition contains XML elements specifying the mapping on attribute level. These XML elements all end with the word *Converter* and start with the type of conversion required. Note that for multiple usages of input attributes there

```

<?xml version="1.0" encoding="UTF-8"?>
<mapping>
  <init>
    <entity name="OutDBEntityName" id="out1" />
  </init>

  <dependencies>
    <dependsOn>
      <entity name="Input1">
        <joinCondition externalID="@XPATH_EXPR" />
      </entity>
    </dependsOn>
  </dependencies>
  ...
</mapping>

```

**Fig. 3.** Mapping specification on entity level

can be multiple conversion specifications. Thus we only need to differentiate between the number of attributes affected on the output side. Consequently there are simple `OneToOneConverters` as well as `OneToManyConverters` as shown in figure 4. Both types in addition exist in a conditional form in case the mappings are only executed (cf. figure 4) based on certain conditions.

**Example 1: Simple Attribute Transformation.** Specification of a simple forwarding of an attribute is shown in figure 4 (src and first target elements). The source of the attribute is specified by an XPATH expression, which is executed on the XML input for the current entity. The value of this attribute is then set as value of the specified output entity and attribute. The entity name references a name defined in the `init` or `update` section of the entity's mapping specification (cf. figure 3).

**Example 2: Conditional to Many Transformation.** The example in figure 4 shows two things in addition to the previous example:

```

<converter name="OneToManyConditionalConverter">
  <condition>
    <key>@XPATH_OF_COND_ATTR</key>
    <valuelist>PL,CL</valuelist>
  </condition>
  <src>@XPATH_OF_SRC_ATTR</src>
  <target>
    <entity>out1</entity>
    <attribute>outattrib1</attribute>
  </target>
  <target>
    <entity>out2</entity>
    <attribute>outattrib2</attribute>
    <function inputFormat="yyyy-MM-dd'T'HH:mm:ss"
      outputFormat="yyyy-MM-dd">convertDateTime
    </function>
  </target>
</converter>

```

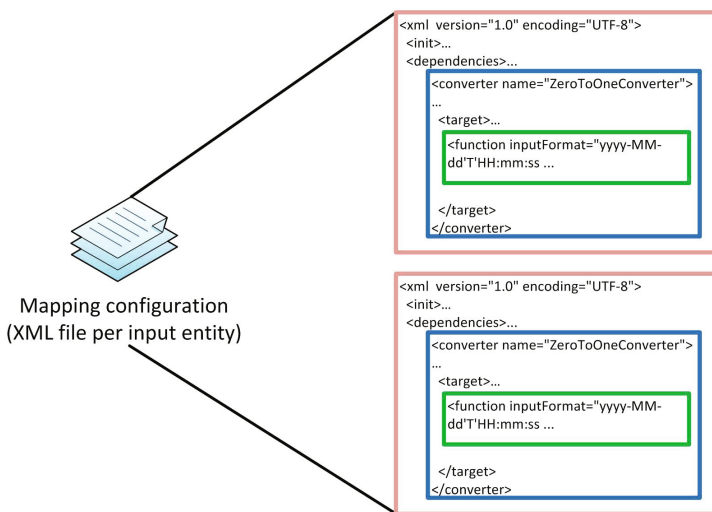
**Fig. 4.** 1:n attribute mapping with conditions and value transformation

- Conditional conversion: converters may be combined with a condition that has to be fulfilled in order for the mapping to be executed. Similar to dependencies at this point only equal conditions are implemented, which check if a certain attribute (of the input, dependent or output elements) is equal to any element of a given list of values. Additional comparisons could be easily added but were not required in our case study.
- To many conversions: an input attribute (or parts of it, see below) may be transformed to different attributes in the output entity. This is achieved by simply adding multiple target clauses to a converter for a single source attribute.

It should be noted that the source attribute to be transformed does not have to be the same as the one on which the condition is evaluated.

**Value Transformations.** The final level of mappings, attribute value conversions, are also specified in the XML mapping specification. As they relate to certain attribute values they can be found inside the `target` element of an attribute mapping as shown for `outattrib2` in figure 4. In this example a datetime value in the input file is transformed to a date value in the output attribute. Such conversion functions can be optionally added to any target attribute specification if required. The type of conversion is specified inside the `function` element and the XML attributes define the details of the conversion. Of course only pre-implemented conversion functions are supported and the semantics of the conversion details depend on the implementation of the function. As explained below in the case study conversions have been implemented in Java allowing for arbitrary complexity on the conversion functions as required.

Figure 5 summarizes how the mappings are specified considering the three different levels of mappings necessary as explained in section 4.1. Mappings on entity level



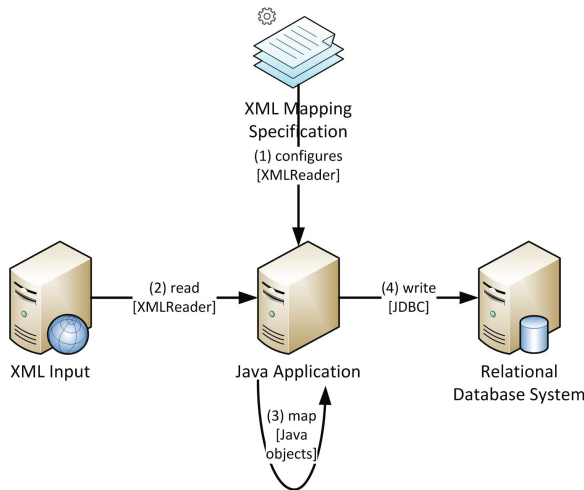
**Fig. 5.** Specification of mappings on different levels



are achieved by an individual mapping specification for each of the input entity types (shown in red). Within these entities attribute mappings are specified by individual converters (blue) whereas transformations on attribute values are defined on the innermost level within the target attribute specification (green). Multiple mappings for individual input objects are achieved by simply multiplying the mapping specifications within the document up to the desired number.

### 4.3 Implementation of Mappings

In the case study an implementation of the whole mapping process in Java has been performed. Technically the system is designed according to the ETL paradigm and it is implemented with different Java processes which together provide the tasks from figure 1. They are combined with a relational DBMS, that also allows for XML data storage (Oracle). Throughout the conversion steps DB transactions are utilized to ensure data consistency. In combination with operating system based fault tolerance (automated process restarts), a high degree of fault tolerance of the overall system is thus achieved.



**Fig. 6.** Implementation of generic mapping system

The Java mapping application dynamically reads the mapping specification documents and checks for syntactical correctness (cf. step 1 in figure 6). Thus changes in the mapping specification simply require a restart of the mapping process without any changes to the source code. After starting the transformation process polls the input folder (or database) for newly arrived XML input (Step 2). Whenever the first input entity of a given type arrives, the mapping specification document is used to instantiate the required converters and functions as Java objects (step 3). The names of the converter XML elements are used to instantiate a corresponding Java class using reflection. This enables dynamic provisioning of converter classes and facilitates complex converters as

the full power of the Java language can be used for implementation. Arguments of the converters and functions are provided to the Java classes based on the specific objects to be processed. Thus additional converters and functions can be easily implemented and added to the mapping processing by observing given interfaces without any changes to the mapping process source code. After successful mapping of an input entity the result is procured to the output system (a relational database in our case, step 4) with standard JDBC operations. Note that while step 1 and the instantiation of mapping objects in the Java application are executed only once, steps 2 to 4 are executed for each input entity with the number of executions depending on the particular mapping specification.

## 5 Conclusion and Outlook

### 5.1 Conclusion

So far the concept and prototypical implementation of eLog have proven to be very promising. The implementation already covers a single exemplary input and also output system. It includes entities that require almost all possible mapping options between input and output systems. The mapping configuration as described previously is defined in an XML file based on a proprietary schema. It may easily be adjusted to different input and output formats (for different versions of the same or other logbooks and maintenance systems). The mapping specification is read dynamically by the mapping component which makes easy and fast adjustments to different versions or products possible without any software development just by configuration. The current output DB resembles the simple relational structure of most airline maintenance systems but may also be adjusted for a different maintenance system.

Both mapping configuration as well as input and output data formats are sufficiently generic in order for the system to be easily adjusted to specific data formats. Thus they might also be used in different domains that also require generic mapping of data between different schemas. The overall modular design of the system by decoupling input and output system leads to a highly scalable overall architecture.

### 5.2 Outlook

By now only a brief evaluation of the first eLog prototype took place. This evaluation mostly included general tests of the system architecture (based on the prototype), some XML database performance tests based on realistic data volume assumptions, tests with different generated flight log data sets, and some stability experiments. All conducted tests showed very promising results for further development.

Future work includes evaluation of eLog in a production environment. Since only selected representative entity types from [1] are used in the prototype, the mapping should be extended to cover the full set. Moreover, additional different aircraft logbook systems and airline maintenance systems should be connected to eLog.

The former will provide additional proof for the scalability and reliability aspects. The latter will lead to the definition of a universal XML schema for the mapping configuration in order to overcome the currently proprietary XML format used. We also consider the examination of ontology based approaches.

In summary automated integration of flight logbooks shows the potential for reduced transfer times of maintenance information coupled with increased correctness when compared to the current manual process. This will eventually lead to reduced maintenance times for aircrafts and thus increase profitability of the airline.

**Acknowledgment.** The authors would like to thank our cooperation partners Lufthansa Technik AG and edatasytems GmbH. The project is part of the Aviation Cluster Hamburg Metropolitan Region's Leading-Edge Cluster Strategy and is sponsored by the German Federal Ministry of Education and Research.

## References

1. Air Transport Association of America, Inc. (ATA). Industry Standard XML Schema for the Exchange of Electronic Logbook Data, 1.2 edn. (May 2008)
2. Conrad, S., Hasselbring, W., Koschel, A., Tritsch, R.: Enterprise Application Integration. Spektrum, Germany (2005)
3. Date, C.J.: An Introduction to Database Systems, 7th edn. Addison-Wesley, U.S.A (2003)
4. Dunkel, J., Eberhart, A., Fischer, S., Kleiner, C., Koschel, A.: Systemarchitekturen für Verteilte Anwendungen. Hanser, Germany (2008)
5. Elmasri, R., Navathe, S.: Fundamentals of Database Systems, 5th edn. Addison-Wesley, U.S.A. (2006)
6. Herr, M., Bath, U., Koschel, A.: Implementation of a Service Oriented Architecture at Deutsche Post MAIL. In (LJ) Zhang, L.-J., Jeckle, M. (eds.) ECOWS 2004. LNCS, vol. 3250, pp. 227–238. Springer, Heidelberg (2004)
7. Hohpe, G., Woolf, B.: Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley, U.S.A. (2003)
8. Inmon, W.H.: Building the Data Warehouse. Wiley, U.S.A. (2005)
9. Jiyi, W.: An Extensible XML Mapping Architecture. In: Proc. 26th Chinese Control Conf., China (2007)
10. Kavelaars, A.T., Bloom, E., Claus, R., et al.: An Extensible XML Mapping Architecture. IEEE Transactions on Aerospace And Electronic Systems 45(1) (2009)
11. Krafzig, D., Banke, K., Slama, D.: Enterprise SOA: Service Oriented Architecture Best Practices. Prentice Hall, U.S.A. (2005)
12. Linticum, D.S.: Enterprise Application Integration. Addison-Wesley, U.S.A. (1999)
13. Hunte, O., Kleiner, C., Koch, U., Koschel, A., Koschel, B., Nitz, S.: Automated generic integration of flight logbook data into aircraft maintenance systems. In: 17th GI/ITG Conference on Communication in Distributed Systems (KiVS 2011). OpenAccess Series in Informatics (OASIs), pp. 201–204. Schloss Dagstuhl, Germany (2011)
14. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
15. Boukottaya, A., Vanoirbeek, C.: Schema matching for transforming structured documents. In: Proc. 2005 ACM Symposium on Document engineering (DocEng 2005), pp. 101–110. ACM, New York (2005)
16. Morishima, A., Okawara, T., Tanaka, J., Ishikawa, K.: SMART: A tool for semantic-driven creation of complex XML mappings. In: Proc. 2005 ACM SIGMOD International Conference on Management of Data (SIGMOD 2005), pp. 909–911. ACM, New York (2005)

# Authoring Processing Chains for Stream-Based Internet Information Retrieval Systems

Philipp Katz<sup>1</sup>, Marius Feldmann<sup>1</sup>, Torsten Lunze<sup>2</sup>,  
Sebastian Sprenger<sup>1</sup>, and Alexander Schill<sup>1</sup>

<sup>1</sup> Technische Universität Dresden, Fakultät Informatik, Deutschland  
`philipp.katz@tu-dresden.de`

<sup>2</sup> Communardo Software GmbH, Dresden, Deutschland  
`torsten.lunze@communardo.de`

**Abstract.** Nowadays, Web-based information systems, such as web feeds or enterprise microblogs produce a seemingly continuous and endless stream of messages. Unfortunately, especially information workers currently experience an information overload. Thus, system support is required that enables a reduction of information load based on an automatic preprocessing of these streams. This paper presents an innovative approach to author IIR (Internet Information Retrieval) processing chains applicable for these stream-based business information systems. It is based on a novel system architecture named Spectre for realising highly scalable systems. Using a dedicated authoring tool, concrete systems can be developed efficiently and adapted to specific requirements. The solution has been validated using a prototypical implementation within a concrete business information system.

**Keywords:** Enterprise 2.0, System Architecture, Information Aggregation, Information Retrieval, Component Architecture, Stream Processing.

## 1 Overall Background and Motivation

*In 2007, for the first time ever, more information was generated in one year than had been produced in the entire previous five thousand years – the period since the invention of writing.* [1]

The biggest part of this information is not published via books and stored in libraries but via the Web. Due to the rise of the Web 2.0, a change in the way of gathering and accessing information took place. It is a very widespread approach, that information is not fetched by users but users are *subscribed* to information sources. This approach is reflected by RSS and Atom web feeds, by microblogs such as Twitter and by activity streams offered e.g. by Facebook. During the last years, this trend has reached the external as well as internal communication means of companies. Nowadays, companies offer various news and social media streams to communicate to customers or deploy internal systems such as enterprise microblogging systems, instant messaging solutions or Wikis to increase the transparency and efficiency of internal communication.

As a result, many employees in such companies have to be subscribed to a variety of information sources generating a potentially huge amount of information. This situation leads to a predicament: On one hand, a specific amount of information is highly relevant for their daily work. On the other hand, in many cases it is impossible to follow all the information flowing to the user.

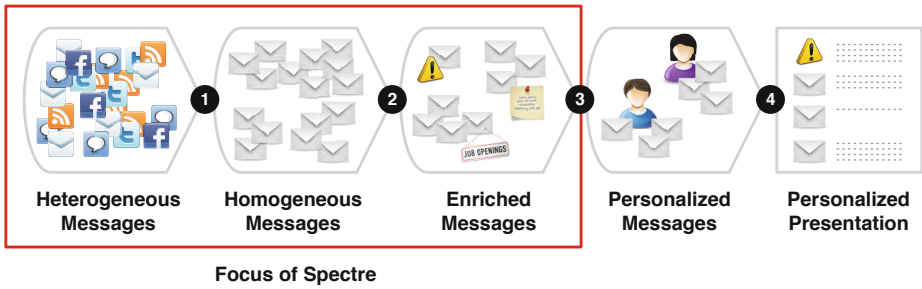
It is obvious, that the problem of information overload is ubiquitous – covering the private as well as and especially the business related usage of current communication means. Thus, system support is absolutely essential to weaken the resulting problem of missing important information. A desired system should process the information automatically and present it in an efficiently accessible manner e.g. by grouping and ranking information or by proposing highly relevant information. Therefore, messages incoming from various stream-based information sources have to be routed through processing chains. A processing chain can be seen as a workflow description with specific IIR preprocessing and processing steps such as tokenization, token filtering, frequency calculation, part of speech (POS) tagging, named entity recognition (NER), text classification, keyword extraction, etc. As a result of our research and development efforts, we have created a novel approach simplifying the overall task from definition to deployment of such processing chains. This paper intends to provide a detailed overview of this approach.

The remainder of this paper is structured as follows. Section 2 discusses a use case to clarify the application area and gives an outline of our approach. In Section 3, an overview of the state-of-the-art is provided. In a further step, Section 4 introduces a highly scalable component-based platform for realizing IIR systems and discusses its characteristics in detail. Based on the information provided in Section 4, the novel authoring approach for developing stream-based IIR systems is discussed in Section 5. Section 6 concludes the paper and gives an outlook on future work.

## 2 Use Case and Approach Overview

The PRISMA information system has already been introduced in [2] and should be familiar to the reader, in order to get a better understanding of the following content. PRISMA forms a novel approach for enterprise communication. It allows its users to subscribe to various heterogeneous information sources as described in Section 1. To help the users to cope with the massive amount of incoming information, data gathered from these sources need to be processed by various IIR tasks that help to rank, group and interconnect gathered data. The illustrated steps take place in Phases 1 and 2 of the processing chain depicted in Figure 1. The system part carrying out the aggregation and preprocessing of incoming messages is named the system's *backend* while the part realizing the user interaction and user management is named *frontend*. The frontend interacts by sending subscription requests to the backend via a well-defined interface.

While the use case presented in [2] describes PRISMA from a end user's perspective interacting with PRISMA's frontend (see Figure 2), the following



**Fig. 1.** Processing Chain (modified version of Figure 1 in [2])

use case will focus on the backend part, and thus the distributed infrastructure and the necessary processing steps for acquiring and processing information. Besides discussing the system architecture, this work will focus on the development methodology of such systems by IIR experts and developers as depicted in Figure 2. For this purpose an easy to use authoring tool is proposed which applies a dedicated authoring methodology discussed in Section 5. It empowers domain experts to specify IIR processing chains without programming skills using existing IIR components.

The following application scenario will be used to illustrate our work: ACME is a fictitious software company with 150 employees using PRISMA [2] to aggregate all relevant information streams. The individual knowledge workers at ACME use the system for consolidating and filtering information from various sources, providing them with a personalised information stream in the system’s frontend with relevant information tailored to their individual interests. It is evident that the computation intensive IIR tasks applied to the messages gathered from the information sources result in problems in regards to the system’s scalability.

After the processing chain of IIR tasks used by a system such as the mentioned one has been defined by IIR experts, it has to be implemented, configured and deployed. As we have made the experience during the development of various IIR systems, such as the mentioned prototype of the PRISMA research project, composing and setting up such a processing chain is highly time consuming and needs a very detailed understanding not only of the IIR domain but especially of building scalable software systems. Besides the huge effort that has to be taken for an initial development of an IIR processing chain, modifications and updates after deploying such a chain are currently labour-intensive. Due to these issues, we developed an overall solution to create and run processing chains within stream-based IIR systems, which consists of two core contributions:

1. A component-based software framework named **Spectre** used to run the processing chain in the form of a highly-scalable distributed system (discussed in Section 4).
2. An authoring tool named **Spectre Cockpit** used to define the processing chains, to configure the applied IIR components and to deploy the chain using the mentioned software framework (discussed in Section 5).

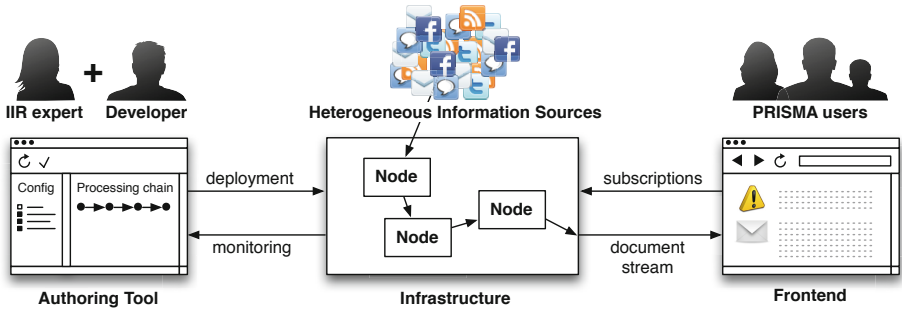


Fig. 2. Overview Spectre Infrastructure

The overall approach is depicted in Figure 2. Using the lightweight framework Spectre, the different IIR tasks are encapsulated into software components which are distributed to a set of available *Nodes*, each of them constituting a virtual or physical machine. The interaction between these components is realized by the Spectre framework as discussed in Section 4. The information determining the execution order of messages by the logic embedded into components and the configuration options for the embedded IIR logic is provided during deployment time by the authoring tool. This information can be updated at any time after the deployment of the components took place. Furthermore, the set of available IIR tasks can be extended, for example when new requirements to the preprocessing results arise.

### 3 State of the Art

After providing an overview of our novel approach, existing systems and solutions for creating processing chains for information streams from various heterogeneous sources are presented. Furthermore, graphical authoring tools for composing information processing and extraction workflows will be presented. Although, NLP toolkits and libraries such as OpenNLP or LingPipe provide basic building blocks, our considerations will focus on architectural aspects enabling a flexible composition and modelling of IIR workflows.

KNIME Desktop is an Eclipse-based platform which offers a component-based workbench for connecting different processing components using a so called pipes and filters approach. It can be used for information integration, data mining and analysis, and machine learning. Though its origins are in the area of bioinformatics, it can also be applied for exploring and experimenting with various data mining tasks in the field of IIR. KNIME Server is a solution to deploy and integrate such KNIME workflows within an existing company's infrastructure by providing the ability to access workflows using SOA paradigms [3]. A similar approach is RapidMine<sup>1</sup>.

<sup>1</sup> <http://rapid-i.com/content/view/181/190/>

UIMA (Unstructured Information Management Architecture)<sup>2</sup> provides an architecture for processing, analyzing and extracting information from unstructured data sources such as text, audio, video and images. UIMA-based systems consist of a number of single components, which can be combined to an analysis workflow. Each component provides an isolated, independent task in the workflow to have a clear separation of concerns, thus allowing a strong modularity, recombining and reusability<sup>4</sup>. UIMA uses a pipeline concept, where each component acts as an Annotator enriching segments in the processed information with specific metadata. UIMA AS (Asynchronous Scaleout) is an architectural approach for achieving scalability for UIMA workflows. Single analysis steps may be replicated and distributed among different logical or physical units. For exchanging data between components, a queue concept is being followed. Data between different nodes is exchanged using Apache ActiveMQ<sup>3</sup> JMS implementation, which provides rudimentary means for monitoring the individual queues' load by using JMX (Java Management Extensions).

AdaptIE<sup>5</sup> provides a language for Information Extraction (IE) tasks allowing to define extraction rules. AdaptIE provides a graphical user interface based on Eclipse focussing on two target groups, supplying each of them with an individual graphical interface: IE experts which are characterized by their fine grained knowledge in e.g. NLP and domain experts who contribute their domain-specific coarse-grained knowledge and build on the IE expert's knowledge.

The MapReduce<sup>6</sup> concept has gained significant attention over the last years, allowing parallel processing of so called big data. However, typical fields of application in the IIR domain are limited to tasks which can be well parallelized, like creating search indices for huge amounts of data, instead of establishing pipeline-based architectures.

None of the aforementioned approaches targets the complete lifecycle of setting up processing chains for IIR tasks. This starts with non-existent solutions which provide the ability to acquire data from various heterogeneous sources of different types of protocols and formats and integrating them into a common processing model with minimal manual effort. While the presented graphical tools such as KNIME Desktop or RapidMiner provide good means for initially prototyping concrete processing workflows, they lack the possibility of transferring such workflows to distributed and scalable infrastructures. Approaches focussing onto the architectural aspect such as UIMA only target parts of the scalability problem, especially do they provide no graphical tool support to set up and extend processing pipelines. Although rudimentary mechanisms for controlling the whole system's load exist, UIMA lacks elaborate techniques for automatically distributing an IIR system's components and to provide dynamic scaling properties of running systems, e.g. by deploying new or additional components to an existing infrastructure if needed.

---

<sup>2</sup> <http://uima.apache.org/>

<sup>3</sup> <http://activemq.apache.org/>



## 4 System Architecture

In this section the architecture of the Spectre framework is discussed. This discussion builds the fundament for providing details about the authoring approach in Section 5. As mentioned in Section 2, the backend offers an interface which allows subscriptions to various types of information sources. Thus, a system frontend sends subscriptions to the backend and receives messages from the associated information source which are then processed and enriched with meta-data by Spectre’s components. In the following, the central characteristics of this scalable backend will be discussed.

The central idea of the architecture is the distribution of processing chains over various nodes. For every subscription (e. g. an RSS feed or a Twitter account), an associated processing chain is determined through which messages resulting from this subscription are routed. Thus, every subscription is associated with one route through available processing components – these routes are named *Channel* in the Spectre terminology. For creating, updating and deleting subscriptions, a centralized system component is responsible which additionally determines an optimal processing chain. With this architecture, the management is centralized while the message streaming is decentralized. Due to this, the architecture is lightweight and easily understandable while highly scalable. The component-based system architecture consists of four core entity types:

1. A set of **Nodes** used as execution environment for Spectre’s components. Every node may be a physical or virtual machine hosting a container in which the Spectre components can be deployed.
2. The system entity named **Coordinator** is responsible for scheduling a processing chain for messages associated with a specific subscription. Furthermore, it serves as point of interaction with a system frontend, thus offering an appropriate interface for subscribing to and unsubscribing from different information sources. Besides these tasks it monitors the overall system state and realizes component replication and distribution to achieve scalability.
3. A set of **Processing Components** per node which offer different IIR-specific functionalities such as stemming, stopword removal, tagging or relation detection. Besides IIR algorithms, logic for accessing information sources is encapsulated in Processing Components, called **Adapter Components**. Spectre offers a predefined set of Adapter Components for example for RSS feeds or Twitter which can be extended easily using Spectre Cockpit.
4. One **Hub** per node that manages the interaction between the components deployed on the associated node in accordance to the schedule provided by the Coordinator. Furthermore, the Hub realizes the interaction with further nodes and transfers messages to these nodes using a queue concept.

Figure 3 and 4 visualize the interaction between these entity types within a concrete example. The overall interaction is divided into three phases. During Phase I, the setup phase, components are initially distributed to nodes, instantiated and registered at the local Hub (Steps 1–3). Furthermore, the available Hubs register at the centralized Coordinator, providing information about the locally

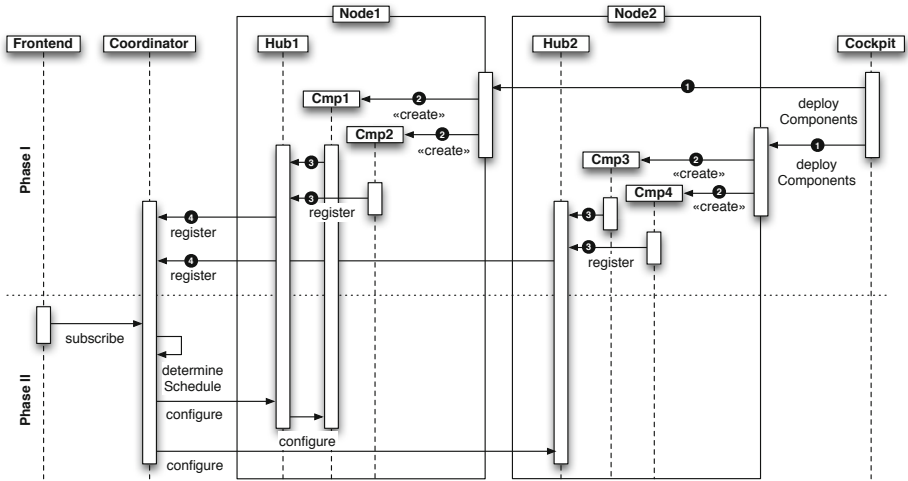


Fig. 3. Setup and Configuration Phase in Spectre

installed components (Step 4). Due to this procedure, every Hub can validate which components are available locally and the Coordinator can verify that all components and Hubs are correctly installed and can participate in processing chains.

After this setup phase, subscriptions can be accepted from a frontend in Phase II, the configuration phase. This phase is started by a frontend transmitting a subscription request to the Coordinator. Every subscription consists of an information source type, such as RSS or Twitter, and further source-specific data, such as authentication information. For every supported type, a specific Adapter Components is offered. It accesses the source and forwards the gathered information to the IIR processing chain. This chain needs to be specified for every subscription. Based on processing information specified by the system author, the processing steps needed for a particular source type are available within the Coordinator. These processing steps determine the order of the component types in the pipeline for each source.

Considering this processing chain, the Coordinator creates an optimal schedule based on the current system state and the capacity utilisation of different processing component instances of the necessary type. In regards to the steps depicted in Figure 4 it is e.g. determined, that incoming messages are processed by the components in the following order: Cmp1, Cmp2, Cmp3 and Cmp4. Due to the components' distribution with ids Cmp1 and Cmp2 to Node1 and with ids Cmp3 and Cmp4 to Node2, the determined schedule is communicated to the appropriate Hubs having the particular components under their local control. This configuration step is repeated for every subscription sent to the Coordinator.

After the configuration information has been distributed for one subscription, messages coming from the associated information source can be routed through

the components. Every incoming message is processed by the respective component and forwarded to the local Hub. Based on the processing chain description received by the Coordinator, the Hub forwards the message either to a further local component or transfers it to a remote Hub. In regards to the presented scenario, the message is forwarded to the Hub of **Node2** after it has been processed by component **Cmp2** so that it can be processed by **Cmp3** next. After having passed the whole processing chain, it is forwarded to the frontend.

During runtime, efficiency and capacity utilisation measures, like throughput and load of components are collected. These values are transferred to the local Hub in a first step. It analyses the collected values and in case, the Hub detects an overloaded component, the Coordinator takes care of rescheduling some of the established channels. The new schedules for the affected subscriptions are distributed to the infrastructure as described in Figure 3. Updates and modifications can be realized seamlessly. By extending or modifying the processing chain or component configuration using Spectre Cockpit, required changes can be transferred to the Coordinator and the participating nodes in the same manner as the initial setup (Phase I) described above. In case of extending processing flows, the additional components are made available ad-hoc to the infrastructure, the affected subscriptions are determined by the Coordinator and the changes are transmitted to the appropriate hubs. Thus, in sum, Spectre allows a transparent modification and extension of processing flows without stopping the processing and with no manual effort except the information provisioning via Spectre Cockpit.

The system architecture has been implemented and technically validated based on OSGI<sup>4</sup> as component execution environment using Apache Felix<sup>5</sup>. To simplify the development of the components, iPOJO<sup>6</sup> has been applied. Furthermore, XMPP<sup>7</sup> was used as a mechanism for inter-Node and Node-to-Coordinator communication. The different components implement a common Java interface which defines the API for controlling the components' lifecycles, for monitoring their workloads and for transferring messages.

## 5 Authoring Approach

After the Spectre framework used for deploying IIR processing chains has been discussed, the associated approach for authoring these chains is presented. As introduced in Section 2, the so called Spectre Cockpit provides an all-embracing authoring tool. It provides IIR experts with the possibility to create and adapt existing workflows when new requirements need to be considered. The authoring process thereby covers the following core areas:

1. New **Adapter Components** (see Section 4) can be created which are then used to acquire data from specific sources and integrate them into Spectre's

<sup>4</sup> Open Services Gateway initiative.

<sup>5</sup> <http://felix.apache.org/site/index.html>

<sup>6</sup> <http://felix.apache.org/site/apache-felix-ipojo.html>

<sup>7</sup> Extensible Messaging and Presence Protocol.

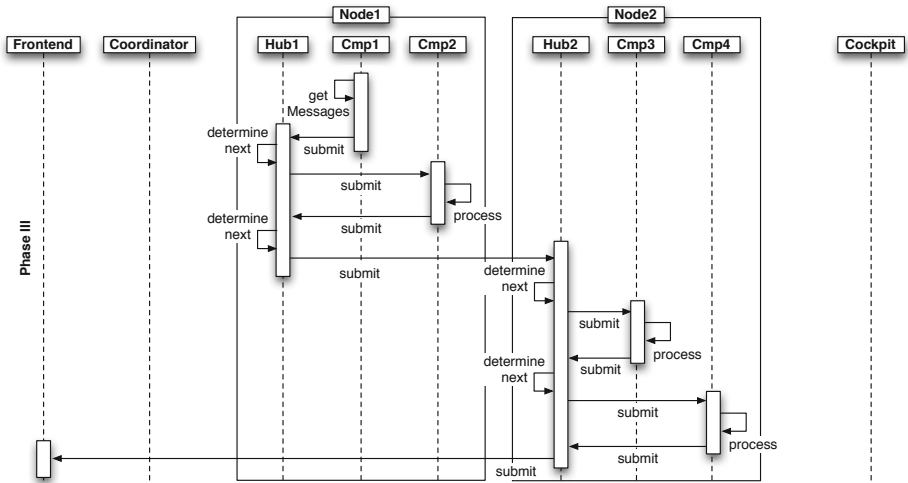


Fig. 4. Processing Phase in Spectre

processing model. As described in Section 4, ready-to-use Adapter Components exist for a range of different source formats. When new types of sources need to be integrated, Spectre Cockpit supports experts with authoring new Adapter Components. This authoring process is realized fully visually without any need for programming. Using a data scheme or message instance from an external information source, the mapping of a source’s data model to the Spectre data model is defined. Furthermore, protocol specific parameters are specified via Spectre Cockpit such as HTTP GET parameters. Based on this visually provided information, Adapter Components can be generated fully automatically.

2. New **Processing Components** for IIR tasks which perform subsequent steps after acquiring source data can be created. These usually carry out typical IIR specific tasks as outlined in Section 1. Similar to the described annotation concepts of state-of-the-art approaches (see Section 3), each Processing Component enriches the documents with specific metadata. Creating Processing Components for IIR tasks is the only step in the authoring methodology for which programming skills are required.
3. Existing components can be configured and combined to **Composite Components** and **Workflows**. Each single Processing Component performs a highly specialised and isolated task. Certain components might thereby require, that specific annotations are already supplied by the incoming document (e.g. a component performing a token filtering which requires an already tokenized document), other components depend on certain configuration parameters (like a language setting for a removal of stopwords). Spectre Cockpit provides the user with an intuitive, workflow-based interface for setting up processing chains and configuring each processing step.

4. **Testing** of components and workflows constitutes an important part of Spectre Cockpit. The authoring tool allows an incremental, experimental approach for creating sophisticated processing chains. Each component in the Workflow can be executed directly within the authoring tool to provide an efficient and error-free development methodology to the IIR expert.
5. After finishing the development process, Spectre Cockpit aids in the **Deployment** of the created workflows. The involved components are distributed to the available nodes and the necessary configuration options are distributed to the Spectre infrastructure.
6. Work on mechanisms for **Monitoring** workflows, nodes and their hosted components is currently in progress. Future versions of Spectre Cockpit will allow to monitor the measures of running systems and allow the user to control and influence strategies for replication and distribution of components.

The components and workflows defined by Spectre Cockpit are transferred to a specific repository from where they can be imported by further Spectre Cockpit instances or shipped to the Spectre execution infrastructure. The authoring tool has been implemented as Eclipse RCP (Rich Client Platform)<sup>8</sup> application, using Eclipse GEF (Graphical Editing Framework)<sup>9</sup> as foundation for the visual workflow editor. The current state of Spectre Cockpit is depicted in Figure 5.

To illustrate the potential of this tool, we build on the given scenario in Section 2. Imagine, CEO Bob decides, that ACME should use the existing PRISMA system to perform opinion mining tasks, monitoring various social media sources for positive and negative user feedback about ACME's products. Mary, responsible as an IT and IIR expert for administrating PRISMA, therefore initially needs to integrate new source types into the processing chain.

Both, Twitter and Facebook are currently not supported by the existing deployment, which means, that new Adapter Components need to be created by using Spectre Cockpit's facilities. Furthermore, Mary requires additional Processing Components for achieving her goal. To detect, which products are mentioned in the processed documents, Mary needs an NER, which is already available as component for Spectre. Additionally, Mary decides to use a text classification algorithm in order to detect the customers' sentiments in the documents. In the described example, such a component is not yet available. Therefore, Mary asks programmer Joe for assistance. Using the Spectre Cockpit, Joe is able to integrate a text classification implementation as new component into Spectre. From his code, a new component which can be integrated into Spectre's pipeline is generated.

After the necessary prerequisites have been fulfilled, Mary can start with creating a new workflow using Spectre Cockpit. It acquires data from Facebook and Twitter using the dedicated components. The incoming data is then piped through various Processing Components like a tokenizer, token remover, etc. After that, a NER is performed by the respective component to detect text occurrences of ACME's products. Using the newly created text classification

<sup>8</sup> [http://wiki.eclipse.org/index.php/Rich\\_Client\\_Platform](http://wiki.eclipse.org/index.php/Rich_Client_Platform)

<sup>9</sup> <http://www.eclipse.org/gef/>

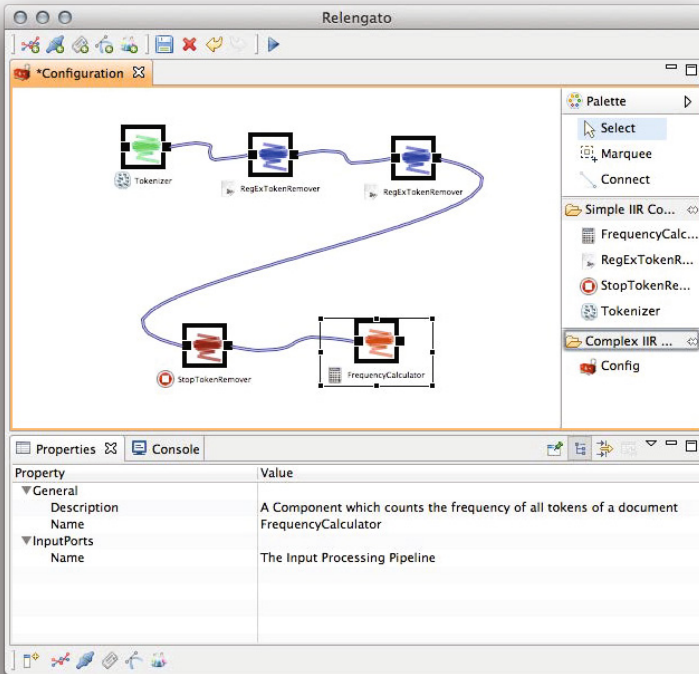


Fig. 5. Processing chain visualized in Spectre Cockpit

component, users' sentiments are detected for each document, assigning meta-data to each document describing the sentiment detection's results.

To summarize, the result of the authoring process consists of the following integral parts: A set of specialized Processing Components performing specific tasks (like the NER and the text classification components in the given example), the configuration options of these components and a workflow description connecting the respective components. This workflow is finally deployed to an existing infrastructure. Spectre takes care of an ad-hoc deployment, distributing the involved components transparently to the available nodes as discussed in Section 4. With this we have provided a detailed overview of the Spectre authoring approach for defining processing chains for stream-based information systems which has been validated by a prototypical implementation applied to real world scenarios in the course of the PRISMA project.

## 6 Summary and Outlook

This paper introduced a novel approach for authoring and executing stream-based IIR systems. The system is based on an easy to apply component-based structure. The core idea is to distribute different IIR tasks using specialized

components over various nodes. Using the dedicated authoring tool Spectre Cockpit, concrete processing flows can be authored and distributed to an infrastructure consisting of various physically distributed nodes. Spectre Cockpit allows efficient modelling of processing workflows for IIR experts by resorting to preexisting Processing Components. Users are isolated from technical aspects, such as deployment and configuration of these components on a sophisticated, distributed architecture. Furthermore, IIR experts are isolated from implementation specific details by providing them an easy-to-use interface for settings up complex processing chains using a visual authoring tool. This paper presented the overall approach. We plan to focus on details of the single aspects and mechanisms used in our approach in several follow up papers.

In future work we will focus on enhancing the scalability of the system by improving the algorithms and means for IIR component distribution and replication. This includes especially the optimization and evaluation of the rescheduling algorithm mentioned in Section 4. Furthermore, an essential aspect in regards of the decentralization of messages will be addressed. The messages are streamed via different ways through the system. Some IIR algorithms need a global perspective on the available data (e.g. for learning models or corpora). It has to be analysed how data or core characteristics of this data can be spread via the components in an efficient manner in order to achieve this global perspective.

Besides these research related issues, it is planned to publish Spectre as an open source project until the third quarter of 2012.

**Acknowledgments.** The contributions presented in this paper are results of the research project PRISMA which is developed in a cooperation between the Technische Universität Dresden and the Communardo Software GmbH. The PRISMA project is funded by the Free State of Saxony and the EU (European Regional Development Fund).

## References

1. Bloem, J., van Doorn, M., Duivestijn, S.: Me the media – Rise of the Conversation Society, VINT edn. (research institute of Sogeti), p. 270 (2009)
2. Katz, P., Lunze, T., Feldmann, M., Röhrborn, D., Schill, A.: System Architecture for Handling the Information Overload in Enterprise Information Aggregation Systems. In: Abramowicz, W. (ed.) BIS 2011. LNBI, vol. 87, pp. 148–159. Springer, Heidelberg (2011)
3. Berthold, M.R., Cebon, N., Dill, F., Gabriel, T.R., Kötter, T., Meinel, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007) (2007)
4. Ferrucci, D., Lally, A.: UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering (2004)
5. Barczyński, W.M., Foester, F., Brauer, F., Schuster, D.: AdaptIE – Using Domain Language concept to enable Domain Experts in Modeling of Information Extraction Plans. In: 12th ICEIS, Funchal, Madeira, Portugal (2010)
6. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI 2004, San Francisco, CA (2004)

# On the Precision of Search Engines: Results from a Controlled Experiment\*

Hasan Girit, Robert Eberhard, Bernd Michelberger, and Bela Mutschler

University of Applied Sciences Ravensburg-Weingarten, Germany  
{giritha, eberharr, michelbe, mutschlb}@hs-weingarten.de

**Abstract.** Handling the growing amount of digital information is one of the major challenges when dealing with the World Wide Web (WWW). In particular, users crave for an effective and efficient retrieval of needed information. In this context, search engines adopt a key role. Besides conventional search engines such as Google, semantic search engines have emerged as an alternative approach in recent years. The quality of search results delivered by search engines is influenced by many criteria. This paper picks up one specific issue, the precision, and investigates and compares the precision of current both conventional (i.e., non-semantic) and semantic search engines based on a controlled experiment with 77 participants. Specifically, Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha are investigated and compared.

**Keywords:** conventional vs. semantic search engines, experiment.

## 1 Introduction

When handling the growing amount of information in the WWW, search engines adopt a key role [1]. The simple use case from a user's perspective: to get an answer (i.e., information) for a specific question (i.e., a search query). However, asking questions (by means of a collection of keywords) and getting suitable answers (by means of relevant search results) remains a big challenge. The reason is that relevant information is indeed typically available, but it remains a complex task to accomplish to identify those information out of the huge amount of available information which are really helpful [2]. Thus, a lot of research is still performed to enable search engines to better answer the questions of their users.

Regarding this simple goal, two major approaches can be distinguished: First, conventional, non-semantic search engines index and rank web pages [3]. When a user enters a search query into a conventional search engine, the engine examines its index (cf. Section 2.1) and provides a list of best-matching web pages according to its internal ranking criteria (which are interpreted by a ranking algorithm). While some conventional search engines, such as Google, index only

---

\* This paper was done in the niPRO research project. The project is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 17102X10. More information can be found at <http://www.nipro-project.org>



selected parts of web pages, others, such as AltaVista, index every single word of every web page [4]. Besides, additional metadata (e.g., author, title, keywords, description, date, language, format) about indexed web pages is used by many conventional search engines as well.

Second, semantic search engines seek to improve search accuracy by understanding user intent and the contextual meaning of terms appearing in the searchable data spaces, whether in the WWW or within closed systems, to generate relevant search results. Rather than using ranking algorithms (such as Google's PageRank) to predict relevancy, semantic search engines use semantics and the science of meaning in language to produce relevant search results. The goal is to deliver the information queried by a user rather than have a user navigate through a list of loosely related keyword results.

Now, which approach is better? This question is difficult to answer. Many issues determine the quality of search results delivered by search engines. This paper picks up one specific issue, the precision, and investigates and compares the precision of both conventional (i.e., non-semantic) and semantic search engines based on a controlled experiment with 77 participants. The investigated search engines include Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha (the reasons for having selected these engines are discussed in Section 3).

This paper is organized as follows. Section 2 provides important background information. Section 3 describes the research design underlying our empirical study. Section 4 presents the experiment results. Section 5 discusses related work. Section 6 concludes with a summary and an outlook.

## 2 Background Information

This section provides background information needed for the further understanding of the paper. Section 2.1 deals with the underlying concepts of both conventional and semantic search engines. Section 2.2 discusses the issue of precision, the key performance indicator we are investigating in our experiment.

### 2.1 Conventional vs. Semantic Search Engines

Conventional search engines gather, index, and rank information [5]. Specifically, these tasks are performed by a *crawler*, an *indexer*, and a *query engine*. Figure 1 illustrates how these three components are applied to process a query [6].

First, crawlers (also named *web spiders* or *robots*) autonomously collect available content, e.g., web pages. Think of a web browser which automatically follows every link on a web page. Doing so, the crawler captures as many web pages as possible. Each gathered web page is stored in a database and then indexed by the indexer [7]. When a user now enters a search query (i.e., a set of keywords) in the search engine's user interface (i.e., the search field), the (inverted) index is combined with a ranking algorithm to generate a list of potential matches, i.e., search results which probably provide relevant information (answers) with respect to the specified search query (question).

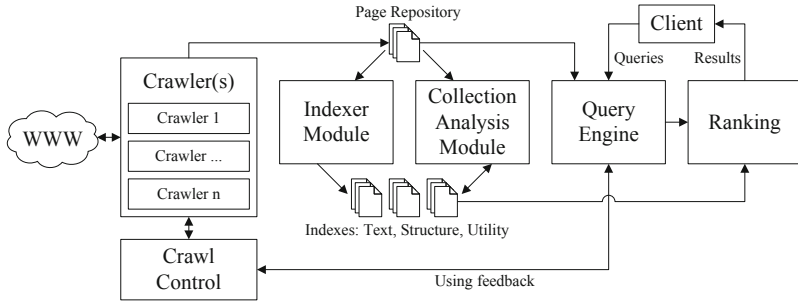


Fig. 1. Conventional search engines

Semantic search engines, in turn, allow users to search not only based on a set of keywords. Natural language search phrases (e.g., when was Google founded?) are used. Moreover, semantic search engines typically allow to further refine the search space in order to increase the accuracy and relevance of search results [8,9]. Generally, there exist three approaches of semantic search engines [10]: *context-based*, *evolutionary*, and *semantic association discovery search engines*. In our experiment we focus on context-based search engines as this approach is used by most existing semantic search engines. Figure 2 illustrates how a context-based semantic search engines generate its results [10].

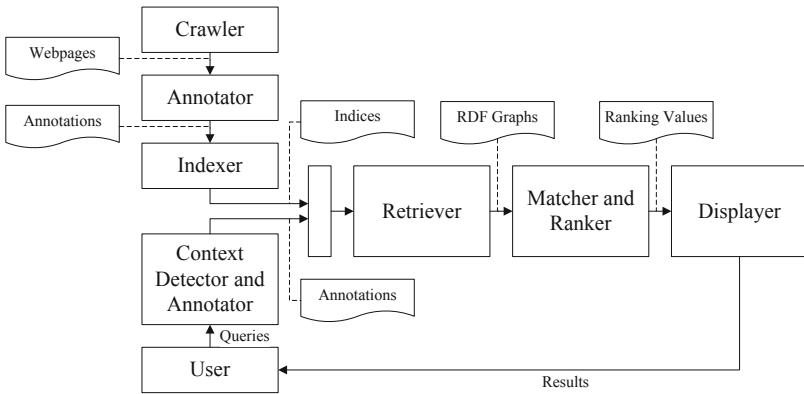


Fig. 2. Semantic search engines

## 2.2 Evaluating Search Engines

As aforementioned, we investigate the precision of search engines in our experiment. This criterion is often used when evaluating a search engine’s effectiveness [11]. The precision describes the ratio of relevant results with respect to the total issued results of a search query, i.e., precision is a measure of the ability

of a search engine to present only relevant results [12]. The most common way to analyze the precision of a search engine is to use a simple binary relevance judgment. A result for a search query is either relevant or not [11]. Let  $a$  be the number of *relevant results* and  $b$  be the number of *non-relevant results*. The precision of a search engine  $p$  can then be calculated as follows [13]:

$$p = \frac{a}{a + b} \quad (1)$$

Though, in order to calculate the precision of a search engine, it becomes obviously necessary to take a closer look at the notion of relevance [14,15]. Crestani and Lalmas define relevance as logical relevance [16]: "A stored sentence is logical relevant to (a representation of) an information need if and only if it is a member of some minimal premiss set of stored sentences for some component statement of that need". We pick up this definition and additionally include two more variables into our definition of precision: Let  $c$  be the number of results containing *links to relevant content* (e.g., a search result is a web directory that is linking to other relevant web pages) and  $d$  be the number of *no results* (e.g., a search result links to a web page which is not reachable). This results in the following adapted equation (2), which we use in our experiment:

$$p = \frac{a + \frac{c}{2}}{a + b + c + d} \quad (2)$$

Note that a search result which contains links to relevant content is more valuable than a non-relevant result, i.e., it must be rated higher. For this purpose, links to relevant content are considered in the numerator of our precision equation. Relevant results ( $a$ ) can be reached within one click (e.g., from the search result to the relevant content), whereas results containing links to relevant content ( $c$ ) need at least two clicks (e.g., from the search result to the link to the relevant content). Therefore,  $a$  is still fully-weighted and  $c$  is half-weighted in the numerator of our precision equation. As a consequence, a search engine has a higher precision when providing links to relevant content instead of non-relevant results. The following section describes the research design underlying our empirical study.

### 3 Experiment Design

The objective of our experiment is to compare the precision of search engines. Therefore, each test person evaluates the relevance of search results for a given search query. Both conventional and semantic search engines are included in the experiment. Doing so, the following experiment variables have to be specified: *search engines*, *search queries*, *test persons*, and *data collection*.

**Search Engines:** First, the search engines to be investigated have to be selected. *Google* as the world's leading search engine has to be considered in any case. Additionally, we selected *AltaVista* as it uses another search algorithm

when compared to Google [1]. AltaVista uses the same search algorithm as Yahoo!. As a third conventional search engine we selected the meta-search engine *MetaGer*. This search engine is actually not a search engine on its own. MetaGer forwards entered search queries to various other search engines and then classifies and ranks the obtained search results [17].

Besides, we included the following semantic search engines in our experiment: *Hakia*, *Kngine*, and *WolframAlpha*. *Hakia* computes search queries both formulated in natural language and collections of keywords. Results are categorized, e.g., in web results, news, tweets or images [18]. *Kngine* is an abbreviation for "knowledge engine". Instead of indexing the web page, *Kngine* tries to interpret the content of web pages and organizes gained information in knowledge databases. It returns both organized, prepared information as well as conventional lists of web pages for a search query. Finally, *WolframAlpha* computes natural language search queries [19]. In a first step, *WolframAlpha* extracts relevant terms of a search query. In a second step, these terms serve as input for internal algorithms (note that almost no information is known on these algorithms). Finally, one (and only one) result is returned for a given search query.

Altogether, we investigate six search engines: two conventional non-semantic engines, one conventional meta-search engine, and three semantic search engines.

**Search Queries:** As explained, we want to investigate the precision of search results in our experiment. In order to compare search results from the six analyzed search engines, we use pre-defined search queries. To make sure that our experiment results are not biased by a too narrow or unfavorable selection of search queries, we use a wide range of topics and search queries. Moreover, we include both *semantic search queries* (which are formulated using natural language) and *non-semantic search queries* (which comprise a set of keywords). Specifically, we define 50 semantic and 50 non-semantic search queries. As the experiment took place in Germany, the search queries are formulated in German. Table 1 shows four exemplary search queries (translated into English).

**Table 1.** Sample of non-semantic and semantic search queries

Non-semantic search queries	Semantic search queries
dollar rate	Who built the Statue of Liberty?
capital of canada	When was Wikipedia founded?

We then entered each of the 100 search queries into each of the analyzed search engines (cf. Fig. 3 - Step 1). The first ten search results delivered by each engine were copied into a separate text document (cf. Fig. 3 - Step 2); hence, six text documents belong to one search query whereas each document comprises the first ten search results of one engine. Doing so, we anonymized and standardized the

presentation of the search results in order to avoid that user ratings are biased by individual preferences, e.g., for certain search engines. During the experiment, the text documents are the basis for evaluating the search results.

**Test Persons:** The overall number of participants in our experiment has to be high enough to ensure that our evaluation results are statistically sound. Literature suggests that at least 50 test persons should participate [20].

**Data Collection:** We used a web-based questionnaire to collect data from the test persons. In order to have different analysis options of the collected data, every test person had to denote some personal data: gender, age, educational background, working position, frequency of internet usage, and frequency of search engine usage. During the experiment, the test persons had to rate each search result as (1) *relevant*, (2) *not relevant*, (3) *links to relevant content*, or (4) *no result*. A *relevant* search result contains useful information for the test person with respect to the search query. Selecting *not relevant* means that the search result has no relation to the search query. If the search result itself does not contain useful information, but links to further relevant information instead, the option *links to relevant content* can be used. In order to handle WolframAlpha (remember that it does only deliver one search result with no outgoing links) the statement *no result* is also added as a possible rating.

When performing the experiment, each test person received an e-mail containing a short description explaining the experiment and its goals. The e-mail includes the prepared text documents (results) (cf. Fig. 3 - Step 3) and a link to the web-based questionnaire (cf. Fig. 3 - Step 4).

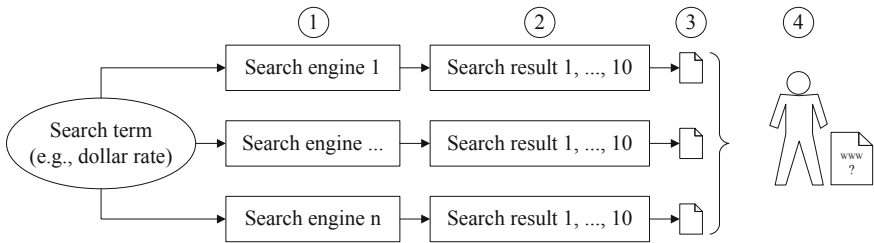


Fig. 3. Performing the experiment - Step 1 - 4

## 4 Experiment Results

In our experiment, 77 people participated. Most participants (58%) were between 16 and 25 years old. Another 34% were between 26 and 35 years old and 4% were between 36 and 45 years old. Only 1% of the participants was between 46 and 55 years old. The rest of the participants (3%) were older than 55.

We also asked for the frequency of internet usage. The majority (61%) told us that they use the internet more than 3 hours a day. Another 35% use it up to 3 hours a day. Only a minority of 4% is online less than 5 hours in a week.

We also wanted to know, how often the participants use search engines. The majority of participants (62%) use search engines more than 3 times a day. Another 25% use respective engines up to 3 times a day. Only 12% use search engines only up to 5 times in a week. The rest of the participants (1%) use search engines less than once in a week.

Moreover, in our experiment the participants evaluated 770 semantic and 770 non-semantic search results for each search engine except for WolframAlpha with only 77 search results. To ensure comparability with the other search engines the results of WolframAlpha were extrapolated (cf. Fig. 4). Table 2 and Table 3 show the raw data collected during the experiment.

**Table 2.** Raw data "non-semantic"

Search engine	relevant	links to	not relevant	no result
Google	362	142	237	29
Hakia	196	112	367	95
AltaVista	228	106	343	93
Kngine	281	118	327	44
MetaGer	192	99	426	53
WolframAlpha	30	0	32	15

**Table 3.** Raw data "semantic"

Search engine	relevant	links to	not relevant	no result
Google	372	110	246	42
Hakia	162	73	410	125
AltaVista	211	70	443	46
Kngine	245	85	385	55
MetaGer	171	84	441	74
WolframAlpha	28	0	30	19

Figure 4 compares the total number of identified relevant results, links to relevant content, non-relevant results and no results for both non-semantic (cf. Fig. 4A) and semantic (cf. Fig. 4B) search queries. Figure 4 shows that Google delivers the best results for both semantic and non-semantic search queries.

Figure 5 additionally shows that relevant results differ between semantic and non-semantic search terms. All investigated search engines except for Google provide a smaller number of relevant results for semantic search queries than for non-semantic ones (cf. Fig. 5A).

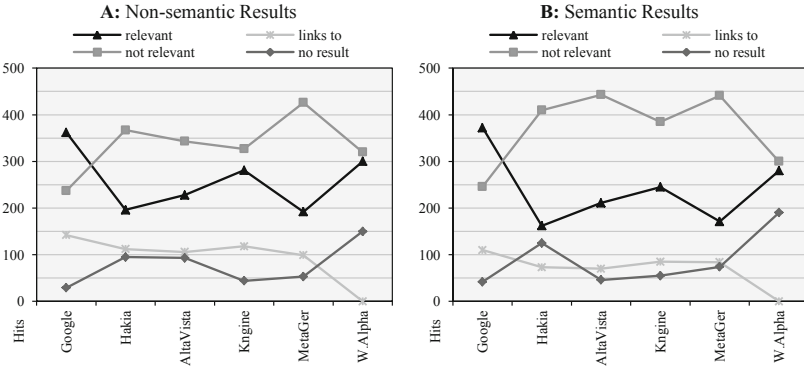


Fig. 4. Comparison I - total number of search results

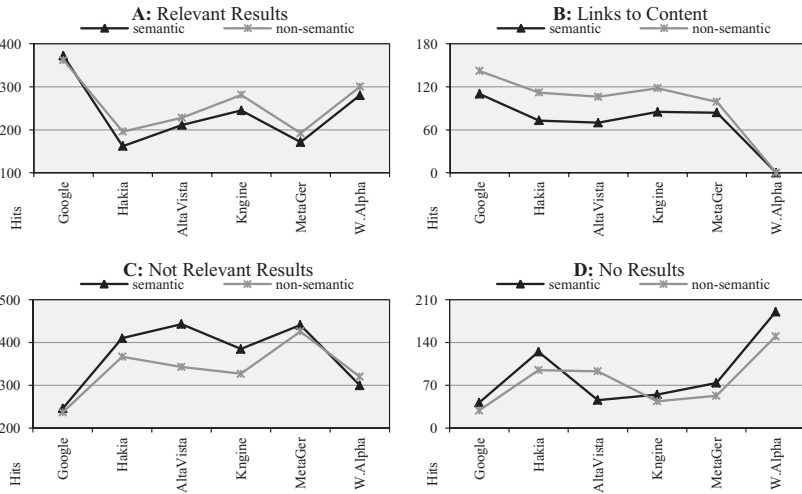


Fig. 5. Comparison II - total number of search results

In order to determine the quality of the investigated search engines in detail, we use equation (2) from Section 2.2. The following results were obtained for the precision (cf. Table 4): Google has a precision of  $p = 0.555$  for semantic search queries and a precision of  $p = 0.562$  for non-semantic search queries. AltaVista has a precision of  $p = 0.319$  for semantic search queries and a precision of

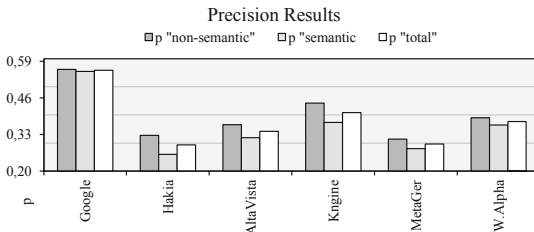
$p = 0.365$  for non-semantic search queries. Kngine, as the best semantic search engine, delivers better results compared to AltaVista with a precision of  $p = 0.373$  for semantic search queries and a precision of  $p = 0.442$  for non-semantic search queries. Table 4 shows the results for all search engines in detail.

**Table 4.** Comparison: Precision "non-semantic" and "semantic"

Search engine	$p$ "non-semantic"	$p$ "semantic"	$p$ "total"
Google	0.562	0.555	0.559
Hakia	0.327	0.260	0.294
AltaVista	0.365	0.319	0.342
Kngine	0.442	0.373	0.408
MetaGer	0.314	0.280	0.297
WolframAlpha	0.390	0.364	0.377

Altogether, all search engines are delivering less relevant search results for semantic search queries - even the semantic search engines. A first reason might be that semantic search queries contain some unnecessary copulas; not all words which must be "understood" to deliver a relevant search result might be identified. A second reason might be that the search engines had problems in handling the German language and the correct interpretation of the (German) search queries. Especially the semantic search engines had significant problems in this respect. These difficulties might lead to search results the user does not consider as relevant. Interestingly, WolframAlpha had the biggest problems.

In summary, best results (cf. Fig. 6) are achieved by Google with an overall precision  $p$  of 0.559 followed by Kngine with a precision  $p$  of 0.408. Third best is AltaVista ( $p = 0.342$ ) followed by MetaGer ( $p = 0.297$ ) and Hakia ( $p = 0.294$ ). WolframAlpha which has to be separately evaluated, has a precision  $p$  of 0.377.



**Fig. 6.** Comparison III - precision of search engines



## 5 Discussion

Our experiment results show very good results for Google when compared to all other search engines. The results indicate that Google is currently indeed the top of the notch search engine. Semantic search engines generally showed relatively poor results. Reason might be, as aforementioned, linguistic deficits in the translation of our queries (which were formulated in German). However, note that all search engines are classified as multilingual, i.e., handling different languages actually should not be a problem.

To better understand the needs of search engine users, we asked the test persons (using the web-based questionnaire) to give feedback on the three most important characteristics of a search engine. Figure 7 shows that the quality and the actuality of results as well as the speed of a search engine regarding the processing of search queries are considered as particularly important.

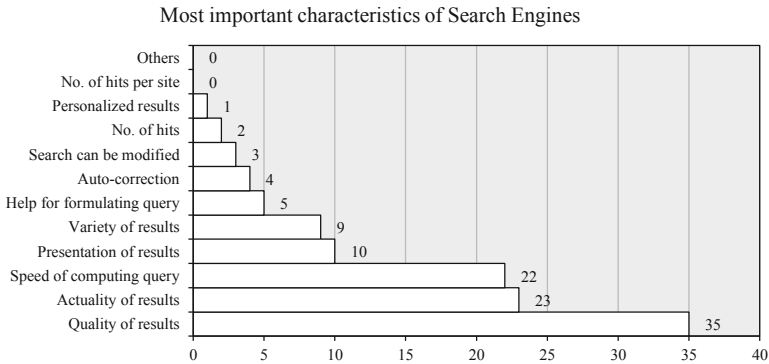


Fig. 7. Most important search engine characteristics

## 6 Related Work

There are other studies dealing with the comparison of conventional and semantic search engines. Empirical and interdisciplinary studies combining semantic web and conventional information retrieval approaches are provided by several authors. Hendler [21], for example, investigates the capabilities of semantic technologies towards their ability to increase the value of content (or search results) through the linking of content. Our experiment, by contrast, only considers retrieved search results (and therewith the precision). The work by Xu [22] provides interesting insights on the impact of collaborative filtering (based on Web 2.0 approaches) on the quality of search results.

Further relevant studies stem from Finin [23] and Ding [24,25]. The relevance of search results of conventional search engines is also investigated by Brin [26], Page [27], and Silverstein [4]. Silverstein [4], for example, analyzed the AltaVista search engine query log comprising approximately one billion entries for search requests over a period of six weeks.

## 7 Summary and Outlook

This paper investigates and compares the precision of both conventional and semantic search engines based on a controlled experiment with 77 participants. The six search engines Google, AltaVista, MetaGer, Hakia, Kngine, and WolframAlpha are investigated. Best results are achieved by Google with an overall precision of  $p = 0.559$  followed by Kngine with an overall precision of  $p = 0.408$ . Third best search engine is AltaVista ( $p = 0.342$ ) followed by MetaGer ( $p = 0.297$ ) and Hakia ( $p = 0.294$ ). WolframAlpha (which is evaluated differently) shows an overall precision of  $p = 0.377$ . In summary, semantic search engines (e.g., Kngine) do not yet achieve the same relevance ratings as conventional search engines (e.g., Google).

Future work will include further controlled experiments in order to evaluate other criteria determining the performance of both conventional and semantic search engines (e.g., further investigation on the recall will be done). Additional research will be also done in the context of enterprise search. Enterprise search engines will be investigated and compared regarding their performance.

## References

1. Levene, M.: An Introduction to Search Engines and Web Navigation, 2nd edn. John Wiley & Sons, Inc., Hoboken (2010)
2. Cambazoglu, B.B., Beaze-Yates, R.: Scalability Challenges in Web Search Engines. In: Book of Advanced Topics in Information Retrieval, pp. 27–49. Springer (2011)
3. Gordon, M., Pathak, P.: Finding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines. *J. Information Processing and Management* 35(2), 141–180 (1999)
4. Silverstein, C., Heinzinger, M., Maires, H., Moricz, M.: Analysis of a Very Large Web Search Engine Query Log. *J. ACM Sigir Forum* 33(1), 6–12 (1999)
5. Risvik, K.M., Michelsen, R.: Search Engines and Web Dynamics. *J. of Computer Networks* 39(3), 289–302 (2002)
6. Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., Raghavan, S.: Searching the Web. *J. of ACM Transactions on Internet Technology (TOIT)* 1(1), 2–43 (2001)
7. Langville, A.N., Meyer, C.D.: *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton Univers. Press, Princeton (2006)
8. Mika, P.: Ontologies are us: A Unified Model of Social Networks and Semantics. *J. of Web Semantics: Science, Services and Agents on the World Wide Web* 5(1), 5–15 (2007)
9. Mayfield, J., Finin, T.: Information Retrieval on the Semantic Web: Integrating Inference and Retrieval. In: Proc. of the Int'l Workshop on the Semantic Web at the 26th Int'l ACM SIGIR Confe. on Research and Development in Information Retrieval, Toronto, Canada (2003)
10. Esmaili, K.S., Abolhassani, H.: A Categorization Scheme for Semantic Web Search Engines. In: Proc. of the 2006 IEEE Int'l Conf. of Computer Systems & Applications, pp. 171–178 (2006)
11. Vaughan, L.: New Measurements for Search Engine Evaluation Proposed and Tested. *J. Information Processing and Management* 40(4), 677–691 (2004)

12. Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. Deutsche Gesellschaft für Informationswissenschaft und Informationsspraxis e.V (DGI), Frankfurt am Main (2005)
13. Tague-Sutcliffe, J.: The Pragmatics of Information Retrieval Experimentation, Revisited. *J. of Information Process Management* 28(4), 467–490 (1992)
14. Mizzaro, S.: How Many Relevances in Information Retrieval. *J. Interacting with Computers* 10(3), 303–320 (1998)
15. Cooper, W.S.: A Definition of Relevance for Information Retrieval. *J. Information Storage and Retrieval* 7(1), 19–37 (1971)
16. Crestani, F., Lalmas, M.: Logic and Uncertainty in Information Retrieval. In: Agosti, M., Crestani, F., Pasi, G. (eds.) *ESSIR 2000. LNCS*, vol. 1980, pp. 179–206. Springer, Heidelberg (2001)
17. Yang, X., Zhang, M.: Necessary Constraints for Fusion Algorithms in Meta Search Engine Systems. In: *Proc. of the Int'l Conf. on Intelligent Technologies*, pp. 409–416 (2000)
18. Campesato, O., Nilson, K.: *Web 2.0 Fundamentals with AJAX, Development Tools, and Mobile Platforms*. Jones and Barlett Publishers LLC (2011)
19. Weikum, G.: Search for Knowledge. In: *Proc. SeCO Workshop on Search Computing Challenges and Directions*, pp. 24–39 (2009)
20. Buckley, C., Voorhees, E.M.: Evaluating Evaluation Measure Stability. In: *SIGIR 2000 Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 33–40 (2000)
21. Hendler, J., Golbeck, J.: Metcalfe's Law, Web 2.0, and the Semantic Web. *J. of Web Semantics: Science, Services and Agents on the World Wide Web* 6(1), 14–20 (2008)
22. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the Semantic Web: Collaborative Tag Suggestions. In: *Proc. of the Collaborative Web Tagging Workshop at the WWW 2006* (2006)
23. Finin, T., Ding, L.: Search Engines for Semantic Knowledge. In: *Proc. of XTech 2006: Building Web 2.0* (2006)
24. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Pen, Y., Reddivari, P., Doshi, V., Sachs, J.: Swoogle: A Search and Metadata Engine for the Semantic Web. In: *Proc. of the 13th Int'l Conference on Information and Knowledge (CIKM 2004)*, pp. 652–659 (2004)
25. Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., Reddivari, P.: Search on the Semantic Web. *J. Computer* 38, 62–69 (2005)
26. Brin, S., Page, L.: The Anatomy of a Large Scale Hypertextual Web Search Engine. *J. Computer Networks ISDN Syst.* 30(1-7), 107–117 (1998)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford InfoLab, Technical Report (1999)

# Towards Semantic Search for Improved Decision-Making

Darijus Straszunskas<sup>1</sup> and Stein L. Tomassen<sup>2</sup>

<sup>1</sup> Straszunskas Forskning, Norway

<sup>2</sup> Microsoft, Norway

contactme@strasunskas.biz, sltomas@gmail.com

**Abstract.** Search applications empower users by simple and quick access to information. However, there is also a significant amount of multi-query sessions originating from a necessity to make a decision, for instance, decide which place to visit. Consequently, the ultimate value of the information provided by a search engine is directly related to the quality of the decision made. Most of current search systems focus on providing a ranked list of information sources relevant to a user's query, while high quality decision-making is characterized by an explicit account on feasible alternatives, reliable information, and clear trade-offs. In this paper we outline our vision on how decision-making can be enabled by enhanced Web search.

**Keywords:** future search system, information retrieval, ontology, decision-making, decision data model.

## 1 Introduction

Search engines have had a tremendous impact on evolution and success of the Web. They have provided easy access to vast amount of information that would have been difficult to access otherwise. However, many existing systems and approaches focus on providing a list of information sources (e.g., links to documents, web pages) relevant to a user's query without accounting for the ultimate goal of the users – to make a decision. In decision-supporting semantic search not only what is retrieved is important but also why it is retrieved and how it is presented with respect to decision in question. This endeavor is not easy – “nothing is more difficult, and therefore more precious, than to be able to decide” Bonaparte Napoleon. In this paper we argue that there are many available technologies that facilitate implementation of this vision as well as skills of end-users have reached quite high sophistication level. On those premises the paper argues that decision-centric search engine is feasible and should be included into research agenda.

Let's consider a futuristic scenario that we will use here to illustrate our vision as follows. A video meeting system facilitates the meeting by procurement and processing of relevant information to have a successful meeting – to come to an agreement or to make a decision. This system automatically retrieves information based on the provided agenda and presents it in an organized manner. Furthermore, the system keeps track of the conversation (i.e. by voice and video) and can, therefore, in combination with the agenda gather, analyze, and present previous statements and viewpoints of the speakers, as well as to simultaneously retrieve

information that supports current topic of discussion. All the gathered data are analyzed and presented, with e.g. statistical data, in an organized manner to support the meeting and decision-making.

Traditional search tools could have been used to gather the information described in the above scenario. However, a high quality decision-making is first characterized by an explicit account on feasible alternatives. Typically, for a high quality decision, extensive information, including both positive and negative aspects (preferably aggregated and grouped accordingly), about the subject at hand is needed. With currently available solutions, this complex information discovery and analysis would require multiple queries on different information repositories, as well as different search strategies. In addition, decision-makers would have needed to perform information analysis, knowledge acquisition, aggregation of data, and qualitative judgments of results [19]. This process is certainly tedious and often suffers from overly narrow view on information and knowledge completeness [18], often referred to as confirmation bias [21], i.e. when information is sought and interpreted to support existing hypothesis and expectations. Consequently, decisions are made without a complete account on alternatives. Ideally, a search system should structure information explicitly based on available alternative actions and guide the decision-maker through key aspects to be considered.

Moreover, in enterprises unstructured information is estimated to represent 80-85% of total information [13, 25] that ought to be used in decision-making and hence needs to be analyzed [22]. Structured data are successfully processed by Business Intelligence (BI) tools. However, current BI tools are not yet ready to integrate unstructured information. Possible convergence of these technologies expresses concern on BI tools becoming more complex [17]. From another hand, availability of technologies for Rich Internet Applications even further enhance Web search experience for users, while semantic technologies provide means to integrate both structured and unstructured data resources and reason over knowledge models [10]. These technologies open up new possibilities for search applications to advance information processing and human-computer interaction.

The described scenario is just one of potentially many use cases of decision supporting information retrieval systems. We propose to investigate search from decision-makers' perspective extending current tendencies of looking barely at search tasks and information needs. Decision data models can be used with domain specific models in combination to filter retrieved information as well as guide users during query specification and modification, or through exploration of results. The objective of our research is to improve the effectiveness of decision-making by use of advanced information integration and analysis techniques, e.g., semantics, knowledge models, text analytics, rich user interfaces, and explore new techniques for knowledge acquisition and presentation. This paper is our first step in that direction, providing discussion on available techniques and possible solution to combine these to enable decision-supporting semantic search.

The paper is organized as follows. In Section 2, we describe the state of the art of decision support systems, relevant tools and techniques. In Section 3 we present the hypotheses, and then we describe the approach for enabling the proposed decision support semantic search system. Finally, in Section 4 we summarize this paper and sketch out future work.

## 2 State of the Art

Many organizations put significant efforts to reorganizing their structures to enable critical business decisions and provide decision-makers with right information to make better decisions faster [6]. Organizations implement decision support systems, text search and analytics, and information management tools. However, these tools still are not well integrated. In this section we provide a brief synopsis of relevant tools and techniques and their state of the art.

### 2.1 Decision Support Systems

Decision Support Systems (DSS) span a wide range of computer-based information systems to support business decision-making activities (e.g., recommendation and expert systems). Web-based Support Systems (WSS) are fairly similar to DSS and a fairly new multidisciplinary research field but focuses on utilizing Web resources in supporting human activities [37]. There are research prototypes exploring use of the argumentation theory to enhance decision support [23], yet Business Intelligence (BI) tools are considered to be one of the prominent types of DSS. Gartner [24] describes BI to be a set of technologies to collect, analyze, and present data used to improve decision-making. The intelligence part of BI is defined to be the ability to discover and explain hidden knowledge within business data [15]. Main advantage of BI is integrating structured data and analyzing it from different views. However, around 80% [25] of enterprise information is stored in semi-structured documents that are still beyond BI reach. Yet there is a noticeable interest from BI vendors to add search capability to their tools and connect to document repositories. Alike, though slowly, enterprise search solution providers are moving towards structured data sources. Analysis of possible convergence of these technologies expresses concern BI tools becoming more complex [17].

Complexity of both implementation and use of BI tools is a big concern in industry – it is reported that only around 50% of BI implementation projects are successful and meet targets [24]. Another issue is requirement of advanced skills to use BI tools that typically results in power/key users at companies responsible for composing queries and assembling analytical reports.

### 2.2 Web Search

Search systems are excellent at connecting different sources of information (in comparison to BI tools operating on databases) and they are made to satisfy a wide range of information needs (e.g., data retrieval, information retrieval, question answering). Recently, information seeking strategies and tasks have been acknowledged as one of the central research topics in Information Retrieval (IR) [32, 36]. There are distinguished many different tasks, for instance, Aula [1] as well as others (cf. [19]) differentiate between three main types of search tasks: fact-finding, exploratory, and comprehensive. Fact-finding is often equated to question answering that requires precise information without further analysis of retrieved information by

user. While exploratory and comprehensive search tasks require cognitive processing and interpretation of returned results [19]. These kinds of search can often be time consuming requiring knowledge acquisition, aggregation of data and qualitative judgments of results by the searcher. In the early phase of such a process searchers can rely/utilize social relations to gather or refine their requirements too. In general, social search systems can be grouped into two categories, social answering systems and social feedback systems [7]. The former utilizes people with expertise to answer specific questions (e.g., Yahoo Answers) while the latter utilizes social data (e.g., Twitter) to rank search results. In a study by Evans and Chi [9] was found that users engaged in social interactions 43% of the time to refine information needs before exploring the Web, and about 59% of the time they shared their findings.

Moreover, Spink et al. [29] report on two studies focusing on multitasking search sessions and found that most Web search sessions consist of two queries of approximately two words. However, a significant part of Web search sessions consists of three or more queries. They conclude that multitasking sessions appear to be growing in Web searching. These sessions are longer and, perhaps, are conducted aiming at a certain decision, searching for different topics that support each other, or seeking for alternatives. A useful feature of a search engine for this kind of sessions would be integration of information across different queries and discovery of insights from the intersection that is currently impossible.

Semantic technologies (i.e. reasoning on ontologies) are apt to discover new knowledge. This feature is very useful in decision supporting search, where discovery of feasible alternatives (that often are beyond knowledge of users) is a critical issue. However, current semantic search tools work on annotated documents resulting in coarse retrieval of information. Nevertheless, some of them categorize results to avoid simple listing of results; reformulate queries and answer questions. For a more extensive overview of semantic search systems the keen reader is referred to [30, 31].

### 2.3 Information Processing and Visualization

Web search typically suffers from information overload retrieving millions of pages. List of retrieved information or rather its length constrains users' further actions [33]. The problem of information overload is two-folded: first, it is difficult to comprehend; second, decision-makers overflowed with information tend to increase self-confidence, though actual decision quality is even reduced (Figure 1), as observed in [9].

One should distinguish between raw information and information that is in a form for cognitive processing by decision-makers. As noted by Shirky [26], "it's not information overload. It's filter failure", i.e. information visualization and presentation of results in search systems still have room for improvement. Different techniques are already available for information presentation, filtering, and organization (e.g., [14, 16]). These techniques can radically enhance graphical user interfaces, improve retrieval processes, and results. There are conducted studies on impact of information representation formats on decision-making behavior; however, most of them are done in the financial sector studying the effect of various representation formats [12].

There are substantial evidences that relevance feedback provides improved search performance (e.g. [19]). Nevertheless, practice shows that searchers are unwilling to make these extra steps (*ibid.*). The users need to be more engaged in the search process, implying more interactive and enthralling user interfaces [27]. Furthermore, the users must be put in control of the search process (or, at least, they should perceive being in control), therefore, trust is important.

Reliable information and clear values are prerequisites for users to trust and use the retrieved information. However, the ultimate goal of a search engine is not limited to bare retrieval of documents being relevant to a given query, but to assist users accomplishing their information tasks like satisfying information needs or decision-making. Therefore, the objective of a search engine is to support decision-makers in finding relevant information, present different views (pros and cons using e.g. sentiment analysis, statistics), guide the user (when information is too scarce, ambiguous, or biases are detected).

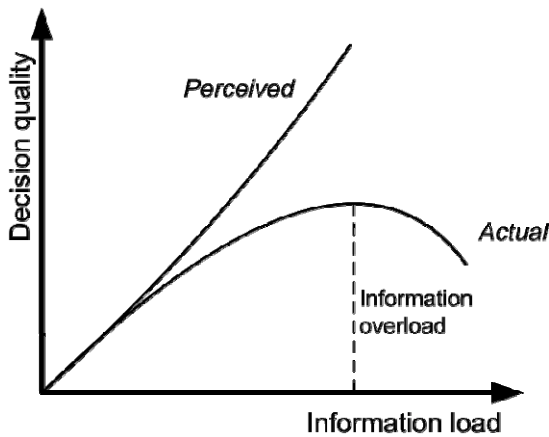


Fig. 1. Information overload and decision quality (adapted from [8])

## 2.4 Text Analytics and Semantics

Text Mining refers to the process of deriving high quality information from both structured and unstructured text [3]. Text mining techniques (e.g. keyword extraction, sentiment analysis, trend detection, clustering, and classification) have been used in industry and academia for some time and are constantly evolving and hence becoming mature. Nevertheless, each individual technique is typically somewhat vulnerable to noise. However, the real power of these techniques lays in the combination of different techniques and hence minimizing their limitations. As a consequent, new and promising techniques are emerging like performing mapping of concepts in concept-space (i.e. ontologies) to words in term-space (i.e. documents) (cf. [31]), construction of tag hierarchies based on folksonomies (cf. [28, 34]), enhancement of classification by sentiment analysis (cf. [35]), or more generally WSS (cf. [37]).

Semantic Web (SW) is a vision of making every piece of information machine-processable and hence enabling a more advanced usage of information elements like



reasoning [2]. Consequently, the SW vision includes many different technologies like Linked open data (LOD). LOD has been given much attention lately as a basis for semantic data processing. LOD refers to a set of techniques and technologies for publishing and connecting structured open data on the Web [4]. One of the basic principles is data being easy to publish and hence easily available. Therefore, a set of guidelines, called the LOD principles, has been proposed (ibid.). Availability of data has increased rapidly, as of 2010 there were more than 27 billion triples (i.e. data elements), an increase by 300% since 2009 [5]. The amount of computer processable data on the Web has overcome a critical mass and hence many potentially useful applications can emerge, currently the availability of applications is lacking.

## 2.5 Summary of Related Technologies

Much of the underlying technology is currently available (e.g., machine learning, knowledge engineering, sentiment analysis, contextual search, concept search, visualization of results). But, to our knowledge, no complete system exists that combines these technologies to support decision support for unstructured texts. Below we summarize two converging technologies:

- Business Intelligence tools:
  - (-) Are complex.
  - (-) Require tedious implementation and sophisticated key users.
  - (+/-) Function on pre-defined knowledge models and on top of structured datasets.
  - (+) Provide powerful analysis and integration of structured data.
- Search systems:
  - (-) Do not account for decision-making needs.
  - (-) Lack of interactiveness with users and aggregation/analytics of so called multitask sessions.
  - (+/-) Empower end-users by simple and quick access to information. However, not all information is reliable, therefore requiring to 'double-check' it before using it for decision-making. This burden is laid upon the shoulders of the decision-maker whom must perform the knowledge acquisition, aggregation of data and qualitative judgments of results.

Obviously, new interaction paradigms and result visualization techniques need to be explored in order to tackle information overload and complexity issues. There is available technology for this - Rich Internet Applications for better user interaction and mash-ups, to integrate, analyze data and textual resources. Text mining techniques and availability of linked open data are highly auspicious for further enhancement of information retrieval.

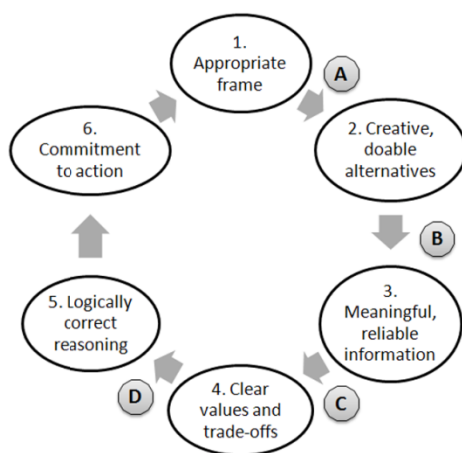
## 3 Proposed Approach

The main goal of the research is to push search technology a step further, i.e. from a query based information retrieval to new insights supporting decision-making. The Web has empowered people by providing an access to the biggest and most vast

information repository as well as providing means to create and share own content. We aim at decision support by making all this information easier digestible and straight forward usable in decision-making process. Therefore, we hypothesize that:

- Many techniques that are necessary to implement decision supporting semantic IR system are already available. They need to be combined, improved or tailored for this specific purpose;
- Users have become more advanced and proficient with the Web systems, especially knowledge workers/ decision-makers;
- There is a huge potential with all this data and information residing on the Web, its usage and utility will be significantly improved by the envisioned approach.

The approach is foreseen to improve information with respect to six requirements of the high quality decision-making cycle [20] displayed in Figure 2, as follows. The cycle starts with a query formulating decision-making need (the 1st requirement). Feasible alternatives (the 2nd requirement) are important for comparison and informativeness in a decision-making. We illustrate a vision of how such system will function by continuing with the video meeting scenario. Let's imagine that the meeting is about how to facilitate and improve information collection and reporting from sales places. The company has many marketing and sales representatives driving around sales points, reporting sales data, interviewing customers, etc. Managers responsible for different geographical regions have gathered in the earlier mentioned video meeting. The future system already has started to collect information (based on agenda) about different solutions, what tools are used for this purpose. After discussing the problem of data collection, analysis, etc., an idea of equipping employees with touch-pads is proposed. Everyone more or less knows what a touch-pad is, the participants engage in conversation about it. The system starts collecting information about different touch-pads, what kind of device they are, what features they have, and, correspondingly, earlier collected information is filtered. Already at this point the system can start supporting decision-making process (see transitions marked by capital letters in Figure 2), as follows.



**Fig. 2.** The six requirements of high quality decision-making (based on [20])

(A) Here, the typical challenge is to find new insights and unknowns to a decision-maker. Based on initial information (i.e. query) and initial relevant results, we foresee the future system trying to infer main features (properties) of an object of interest. In the illustrative scenario, the system finds a set of alternative touch-pads, and displays this information for the participants. To do that, text mining of query results will need to return the light-weight (domain) model representing the object of interest.

(B) To provide meaningful and reliable data (i.e. to satisfy the 3rd requirement), we foresee information structuring and presenting based on the key features (properties). The retrieved list of alternatives will then be used to structure information accordingly. In the example, the touch-pads are grouped based on WiFi and 3G/4G support, memory and screen size. Rich (though intuitive) user interface allows viewing information based on key features inferred.

(C) Pros and cons are further analyzed by finding and displaying facts, opinions. In this example system highlights weakness of touch-screens when operating in low temperatures. To verify such information user may opt to use their social network to try to access first-hands experience, or alternatively, to find a competent service provider to test selected models.

(D) Discussion of use cases, benefits and doubts is simultaneously transcribed and related to the discussed tablets, their features, and ranking them accordingly. This step wraps-up the meeting and finalizes the decision-making support process from the system’s perspective.

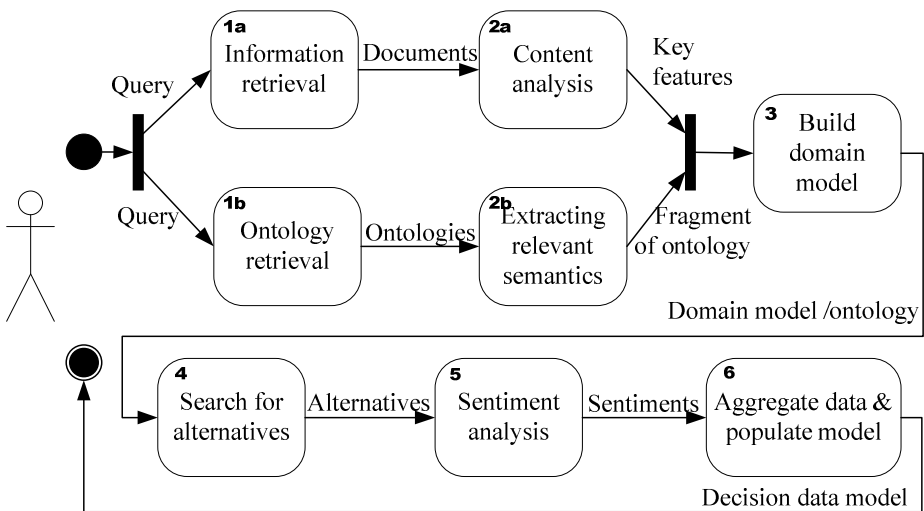


Fig. 3. Logical steps of the approach

To achieve this, from a technical perspective, we foresee the following coarse process of six logical steps as displayed in Figure 3. The first step deals with information retrieval (1a) and retrieval of an ontology (1b) based on a query provided by a user.

Here APIs of existing search engines are used. Then content analysis (2a) is performed to extract key concepts and their characteristics (e.g., using a method described in [31]), and if several ontologies are available - they are compared and a relevant fragment is extracted (2b). The purpose of these steps is to build domain model (3), containing an object of interest, key properties (e.g., functionality, application areas) that are later used to query for alternatives (4). Additional information is gathered and synthesized using sentiment analysis (5) (using a method described in [35]) and, finally, the domain model is populated by aggregated data to produce a decision data model (6).

The decision data model (DDM) is a model containing information about the *object of interest*, its *alternatives*, *key features*, *application areas* and, most important, a *feasibility scores* that aggregate sentiments and are used to quantify probability of usefulness. The DDM is directly used in decision-making as it visualizes the vital information as well as allows exploring the source data. To further support decision-making by probability reasoning and finding the best alternative the DDM is converted to influence diagram (Bayesian network), e.g. [11].

## 4 Conclusions and Future Work

The objective of this paper is to present a vision of a search system for decision-making and to call for research efforts in this direction. We argue that extension of existing search systems towards decision support will even further improve usefulness of the Web. A possible approach of such system is sketched out and a brief overview of existing underlying techniques is provided in the paper. We argue that recent advancements of Web technologies and sophistication of users makes implementation of decision supporting semantic search feasible and needful.

Though this research is still in its infancy, we hope to have revealed its potential and have intrigued the research community to pursue this multi-disciplinary vision. However, to achieve this ambitious goal, the future research needs to answer two broad questions: a) how can Web search tools be integrated and tuned to facilitate a decision-making in a computationally efficient manner?; b) how to effectively and automatically display information that is not necessarily understood by users?

## References

1. Aula, A.: Query Formulation in Web Information Search. In: Isaias, P., Karmakar, N. (eds.) Proc. of the IADIS Int. Conf. WWW/Internet 2003, pp. 403–410. IADIS Press, Algarve (2003)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. *Scientific American* 285(5), 28–37 (2001)
3. Berry, M.W., Kogan, J.: *Text Mining: Applications and Theory*. Wiley (2010)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *Int. J. on Semantic Web and Information Systems* 5(3), 26 (2009)

5. Bizer, C., Heath, T., Berners-Lee, T.: 4th Linked Data on the Web Workshop (LDOW 2011) (2011), <http://events.linkedata.org/ldow2011/slides/ldow2011-slides-intro.pdf> (accessed December 1, 2011)
6. Blenko, M.W., Mankins, M.C., Rogers, P.: The Decision-Driven Organization. *Harvard Business Review* 88(6), 54–62 (2010)
7. Chi, E.H.: Information seeking can be social. *Computer* 42(3), 42–46 (2009)
8. Eppler, M., Mengis, J.: The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society* 20(5), 325–344 (2004)
9. Evans, B.M., Chi, E.H.: An elaborated model of social search. *Information Processing & Management* 46(6), 656–678 (2010)
10. Fensel, D., van Harmelen, F.: Unifying reasoning and search to Web scale. *IEEE Internet Computing*, 94–96 (2007)
11. Fenz, S.: An ontology-based approach for constructing Bayesian network. *Data & Knowledge Engineering* 73, 73–88 (2012)
12. Ghani, E.K., Laswad, F., Tooley, S., Jusoff, K.: The role of presentation format on decision-makers' behaviour. *International Business Research Journal* 2(1), 183–195 (2009)
13. Grimes, S.: Unstructured Data and the 80 Percent Rule (2008), <http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551> (accessed January 13, 2012)
14. Haas, K., Mika, P., Tarjan, P., Blanco, R.: Enhanced Results for Web Search. In: *Proceedings of SIGIR 2011*, pp. 725–734. ACM (2011)
15. Herschel, R.T., Jones, N.E.: Knowledge management and business intelligence: the importance of integration. *Journal of Knowledge Management* 9(4), 45–55 (2005)
16. Hoeber, O., Yang, X.D.: HotMap: Supporting visual explorations of Web search results. *Journal of the American Society for Information Science and Technology* 60(1), 90–110 (2009)
17. Jez, V.: Convergence, Complementarity or Disruption: Enterprise Search and Business Intelligence. Master thesis, BI Norwegian School of Management, Oslo, Norway (2009)
18. Larrick, R.P.: Broaden the Decision Frame to Make Effective Decisions. In: Locke, E.A. (ed.) *Handbook of Principles of Organizational Behavior*. Wiley, Chichester (2009)
19. Marchionini, G.: Exploratory search: from finding to understanding. *Communications of the ACM* 49(4), 41–46 (2006)
20. Matheson, D., Matheson, J.E.: Describing and Valuing Interventions That Observe or Control Decision Situations. *Decision Analysis* 2(3), 165–181 (2005)
21. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomena in many guises. *Review of General Psychology* 2(2), 175–220 (1998)
22. Norheim, D., Fjellheim, R.: AKSIO – Active knowledge management in the petroleum industry. In: *Proceedings of the ESWC 2006 Industry Forum* (2006)
23. Ouerdane, W., Maudet, N., Tsoukias, A.: Argumentation Theory and Decision Aiding. In: Ehr Gott, M., et al. (eds.) *Trends in Multiple Criteria Decision Analysis*, pp. 177–208. Springer, Berlin (2010)
24. Sallam, R.L., Richardson, J., Hagerty, J., Hostmann, B.: Magic Quadrant for Business Intelligence Platforms (2011), <http://www.gartner.com/technology/media-products/reprints/microsoft/vol2/article15/article15.html> (accessed January 13, 2012)
25. Shilakes, C.C., Tylman, J.: Enterprise Information Portals. Merrill Lynch & Co. (1998)
26. Shirky, C.: It's not information overload. It's filter failure (2008), <http://web2expo.blip.tv/file/1277460> (accessed June 1, 2011)

27. Shneiderman, B., Plaisant, C.: *Designing the user interface*. Pearson Addison Wesley (2005)
28. Solskinnsbakk, G., Gulla, J.A.: A Hybrid Approach to Constructing Tag Hierarchies. In: Meersman, R., Dillon, T., Herrero, P. (eds.) *OTM 2010, Part II*. LNCS, vol. 6427, pp. 975–982. Springer, Heidelberg (2010)
29. Spink, A.H., Park, M., Jansen, B.J., Pedersen, J.: Multitasking during Web search sessions. *Information Processing and Management* 42(1), 264–275 (2006)
30. Strasunskas, D., Tomassen, S.L.: On Variety of Semantic Search Systems and Their Evaluation Methods. In: *Proc. of the Int. Conf. on Information Management and Evaluation (ICIME)*, pp. 380–387. Academic Conferences International, Cape Town (2010)
31. Tomassen, S.L.: *Conceptual Ontology Enrichment for Web Information Retrieval*. PhD thesis, NTNU, Trondheim, Norway (2011)
32. Vakkari, P.: Task-based information searching. *Annual Review of Information Science and Technology* 37(1), 413–464 (2003)
33. Venkatsubramanian, S., Hill, T.R.: An empirical investigation into the effects of web search characteristics on decisions associated with impression formation. *Inf. Syst. Front.* 12, 579–593 (2010)
34. Veres, C., Johansen, K., Opdahl, A.L.: Browsing and visualizing semantically enriched information resources. In: *Proceedings of the International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, pp. 968–974. IEEE Computer Society (2010)
35. Wei, W., Gulla, J.A., Fu, Z.: Enhancing Negation-Aware Sentiment Classification on Product Reviews via Multi-Unigram Feature Generation. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) *ICIC 2010*. LNCS, vol. 6215, pp. 380–391. Springer, Heidelberg (2010)
36. Xie, I.: Dimensions of tasks: influences on information-seeking and retrieving process. *Journal of Documentation* 65(3), 339–366 (2009)
37. Yao, J.T.: An Introduction to Web-based Support Systems. *Journal of Intelligent Systems* 17(1), 267–281 (2008)

# Synthetic History for Exchange Traded Funds

Aistis Raudys, Lukas Sirvydis, and Karol Lisovskij

Faculty of Mathematics and Informatics  
Department of Informatics, Vilnius University  
Naugarduko 24, Vilnius  
Lithuania LT-03225  
aistis@raudys.com

**Abstract.** To make money in trading one ought to forecast the future price, but to do so accurately one must verify the predictions using the past data. A short trading history can present a problem. We showed both theoretically and experimentally that the history of some financial assets can be reconstructed quite accurately. We forecasted the past price movements of exchange traded funds (ETFs). The problem in practice is very acute as there are a number of very liquid ETFs that can be traded with minimum slippage but their available history is too short. In such situations systematic traders cannot test their trading models as the history length is insufficient. To forecast historical ETF prices we used stocks with a longer history available. In some cases we created multiple model instances with a variable number of stocks. As soon as the stock history became unavailable we selected a different model. We compared this and eight other methods using a set of US ETFs ranging from S&P 500 to uranium. The experimental study showed the expectation maximisation with covariance matrix normalization to be the best method for this task.

**Keywords:** synthetic history, artificial history, time series, regression, expectation maximisation, imputation, missing data.

## 1 Introduction

Recently, exchange traded funds (ETFs) have become very popular and liquid (cheap to trade/low transaction cost). Traders are always seeking to diversify their trading portfolio and to invest in uncorrelated assets. It is a perfect opportunity to invest in previously inaccessible uncorrelated resources like uranium, lithium or emerging market stock indexes. The downside, however, is that many liquid ETFs have a very limited history. For example, the Global X Uranium ETF (symbol: URA) started trading only on 4<sup>th</sup> November 2010.

Systematic trading is a type of trading that uses computer models to make a trading decision. It is becoming increasingly popular nowadays [1]. Systematic trading uses the past price movement patterns to create profitable trading systems. Some use only market data and some include fundamental factors; all, however, use historical data to test the profitability of their trading. There are a number of investment funds that employ such a trading style. Trend followers are probably the best known among

them (f.e. Transtrend, Bluetrend, Aspect Diversified, etc.). Systematic trading heavily relies on historical trading strategy simulations. If no data are available the profitability of the trading system cannot be tested and no trading can be carried out. Therefore, the topic of this study is very important for this type of business.

In statistics, missing data reconstruction is usually refereed as imputation. It is a very common task to fill gaps in the data and is widely used in social and earth sciences. In social sciences it is very common to fill gaps in surveys [2] and in meteorology it is quite common to fill empty grid items for temperature and other measurements [3]. For more information on various imputation techniques and application areas refer to [4]. In financial time series data is usually recorded better and is rarely missing.

In pattern recognition there are plenty of methods that can be used to fill missing values in the data sets: regression, regression trees, nearest-neighbour, neural networks, expectation maximization (EM), etc. Different methods are needed for different types of missing data. For instance, Shin-Mu Tseng and Kuo-Ho Wang [5] proposed a new method for reconstructing missing values in data where cluster properties exist among the data records. Their work objective was to integrate the clustering and regression techniques for estimating the missing values. In [6] missing data reconstructions were made for monthly river records. Expectation maximization method successfully estimated the missing river flow data. Authors in [7] tried to use decision tree classifier for time series reconstruction where environmental data is recorded by monitoring stations. Moreover decision tree classifier is compared with two signal reconstruction methods: mean value and polynomial interpolation and was offered to use as it outperformed the other methods. In the following paper [8] authors introduced a new approach for reconstructing missing daily precipitation data taken from 39 weather stations. Combination of regression trees and artificial neural networks were used. Xiaolu Huang [9] introduced another method for reconstructing missing data: pseudo-nearest-neighbour, on a Gaussian distributed data sets. This method was compared with constant value substitution and the missing data ignorance (non-substitution) methods. Pseudo-nearest-neighbour method showed the best results to Gaussian data sets among the substitution and non-substitution methods.

Number of papers tried to forecast the future. Unfortunately there are only a few publications addressing the past historical data. Literature on the synthetic time series history creation is very sparse. The reason is obvious. Successful prediction of the future can lead to a successful investment strategy and profit. In this paper, contrary, we try to predict the past. Here, we make a hypothesis that ETF can be reconstructed with a reasonable accuracy from a set of other factors (stocks). In reality it may not be possible to reconstruct the series if quality source is not available. So we also try to define a threshold that ETF is predicted accurately.

## 2 History Reconstruction Methods

The standard prediction method is a linear regression where we try to predict variable  $x_0$ , according to the known values of  $p$  prediction variables,  $x_1, x_2, \dots, x_p$ :  $x_0 = w_1x_1 + w_2x_2 + \dots + w_px_p + w_0$ . To find a vector of weights  $\mathbf{w} = (w_0, w_1, w_2, \dots, w_p)$  most often one uses the standard mean square regression. As described in [10] one may use sample (empirical data) estimates of mean values of variables  $x_0, x_1, x_2, \dots$ ,



$x_p$ , and  $(p+1) \times (p+1)$  dimensional covariance matrix. To estimate the mean vector and the covariance matrix one needs to have empirical data, consisting on  $M$   $(p+1)$  – dimensional observation vectors (training data). The minimal prediction accuracy (mean square prediction error),  $\sigma_{\min}$  depends on true values of  $p+1$  variances and  $p(p-1)/2$  correlations that compose the covariance matrix. This dependency is individual (specific) for each particular prediction task. In general, the more prediction variables,  $x_1, x_2, \dots, x_p$ , we have the smaller is value  $\sigma_{\min}$ .

In practice, however, we use sample estimates. The sample based estimates are not exact. The more training vectors we have, the more exact prediction we obtain. Statistical theory gives a relationship between actual prediction error,  $\sigma_{\text{prediction}}$ , minimal error,  $\sigma_{\min}$ , sample size,  $M$ , and dimensionality,  $p$ ,

$$\sigma_{\text{prediction}} = \sigma_{\min}(p) \times \sqrt{1 + \frac{p}{M - p}} \tag{1}$$

In Equation (1) we use expression  $\sigma_{\min}(p)$  to stress that ideal prediction accuracy depends on a number and quality of prediction variables,  $x_1, x_2, \dots, x_p$ . Equation (1) also indicates that the accuracy increases with an increase in the size of the training data.

Unfortunately, in practical tasks, such as, the exchange traded funds analysis, some very important training data series are short. Therefore, the training data size is bounded by the length of the shortest time series. Our task is to use additional information contained in longer time series in order to increase the length of the training data. To realize this task we are obliged to create synthetic time series. Equation (1) shows that if the history of the reconstruction task would be solved successfully, we could increase the time series prediction accuracy notably in small sample situations where sample size,  $M$ , exceeds dimensionality,  $p$ , negligibly.

For the ETF past history reconstruction, we use various statistical and pattern recognition methods with some variations. We have two groups of methods. One uses the same model for the entire series while the other group uses the different model for the different periods of the history. Variable number of factors influences the prediction at a different point in the history. Our methods reconstruct the time series form a set of factors and report accuracy obtained, then one can choose to use it if accuracy is sufficient.

Let's define a problem:  $x_b, i=q \dots z$  – known values,  $x_l$  – the most recent value, and  $x_z$  – the oldest known value.  $x_b, i=z+1 \dots n$  – unknown values (the values we are seeking to reconstruct),  $z$  – the oldest known value position of the series we are reconstructing,  $n$  – total number of values in the source dataset,  $s_{ij}$  –  $i$ 'th value of the  $j$ 'th source. We try to predict values  $x_b, i = z+1 \dots n$ .

## 2.1 Wiener Process

In this paper, Wiener process is used for creating random history drawn from the normal distribution with the same mean and standard deviation as the existing history set. I.e. the mean and the standard deviation are the same:  $x_i \in N(\mu, \sigma), i=z+1 \dots n, x_i, i \geq z+1$ , and  $\mu$  is the mean of the known sample and  $\sigma$  is the standard deviation of the

known sample ( $x_i, i=1...z$ ). Wiener method is widely used in data generating like stochastic generation of hourly mean wind speed data [11], rainfall data generation model for climate change conditions [12].

### 2.2 The Proxy Method

The Proxy method is the simplest way of solving the problem. It is based on a substitution of one time series with another. We added modification of the standard deviation  $x_i = s_{ij}w, i = z + 1...n$ . Here  $w$  is the standard deviation adjustment weight and  $s_{ij}$  is the most correlated source series. The method is similar to a “Cold – deck” method [13] that replaces the missing values with the constant values from the external source. The only modification is standard deviation adjustment. The example of “Cold – deck” method is presented in a short-term economic survey of all enterprises in Japan [14].

### 2.3 Multiple Proxy Method

Multiple proxy method is based on the proxy method with the modification which extends existing series history using several assets for different periods. The graphical explanation about how multiple proxy method can be used to extend time series history is described in Fig. 1. The history can be extended to the same length as the length of the most correlated asset (see Equation 2).

$$x_i = s_{ij}w, i = n_1...n_2, x_i = s_{ij}w, i = n_2 + 1...n_3, n_i \geq z + 1, \forall i. \tag{2}$$

We select the most correlated assets and sort them in descending order that the first has the highest correlation coefficient and the last is with the lowest correlation. We take the most correlated series ( $s_1$ ) and use that series as a proxy for the first period ( $p_1$ ). If the available series is no longer available we select the next available most correlated series ( $s_3$ ). If some of the series are shorter ( $s_2$ ) and ( $s_4$ ), those series are not used. The same process continues until we exhaust all available series ( $s_n$ ). Dotted vertical lines represent periods which are added to available data history. This illustration has 3 periods added by  $s_1, s_3$  and  $s_n$  time series.

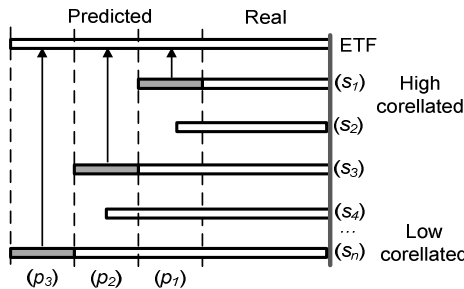


Fig. 1. Multiple proxy method’s graphical representation

We were unable to find similar research on this method, but there exists research works where correlation analysis is used in fuzzy time series prediction [15].

### 2.4 Regression Method

Multiple linear regression analysis is a common statistical method that is used to model the relationship between the dependent variable  $Y$  and independent variable  $X$ . The regression equation:  $y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$ ,  $i=1 \dots N$ . Here,  $\beta$  is the regression model parameter  $\beta = (\beta_1, \dots, \beta_p)$ , and  $\varepsilon_i$  is a random noise. The relationship is outlined by the linear function. The linear function parameters are estimated from the known variables  $Y$  and  $X$ . The estimated parameters and the linear function are called statistical model.

Linear regression models are commonly used in various practical applications. Regression analysis helps to solve air pollution problems [16], predicting the air pollution concentrations in various environments. Certainly, regression method is commonly used in finance. Authors in [17] predict the price movements of the financial security or the index of securities.

#### Independent Variables Preparation for the Regression Model

Regression method extends the ETF history using the most correlated stocks. That purpose is achieved by selecting not just one stock, but a basket of stocks. History lengths of the stocks are different. But regression method requires the same history length (see Fig. 2a).

Below we describe 3 regression model methods that use different length of the stock histories. Suppose that, all method results are of the same length.

1. Choose the longest stock history (see Fig. 2b) from the bucket and create a regression model with one period ( $p_1$ ) and use only that one.
2. Extend all stock histories till the longest stock history length (see Fig. 2c) by filling the empty places with zeros. One regression model is used for the whole period ( $p_1$ ).
3. Create multiple models for various parts of the history depending on the source data availability (see Fig. 2d). For recent history one will use more complex (with more sources) model ( $p_1$ ). For older history, one will use simple models with fewer sources ( $p_2$ ) and ( $p_3$ ).

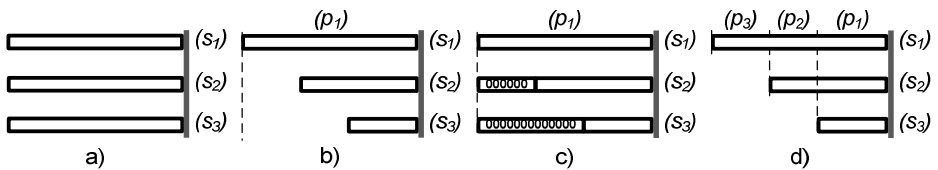


Fig. 2. Different data preparations and estimation of the regression model

### 2.5 Nearest Neighbour Estimator

The  $k$ -nearest neighbour estimator is attractive with its simplicity and the capabilities to solve complex nonlinear problems. In our work the method used to predict the instrument price by selecting the  $k$  similar prices from the same instrument history (see Fig. 3). All these  $k$  prices are averaged and results serve as a prediction for the unknown value.

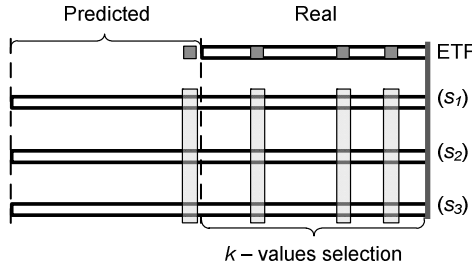


Fig. 3. Nearest neighbour  $k$  – values selection

We find one nearest neighbour for values  $s_{ij}, j=1 \dots 5, s_{lj}, j=1 \dots 5$  and  $l \leq z$ . We use Euclidian distance (see Equation 3).

$$l = \arg \min_{l=1 \dots z} \left( \frac{1}{k} \sqrt{\sum_{j=1}^5 (s_{ij} - s_{lj})^2} \right) \tag{3}$$

We repeat the process and find multiple  $l: l_i, i=1:k$ , later, we average the results to create our prognosis (see Equation 4).

$$x_i = \frac{1}{k} \sum_{u=1}^k x_{i_u}, i \in (z+1 \dots n) \tag{4}$$

Selection of  $k$  values is based on calculation of the Euclidean distance as similarity metric. This method is based on finding similar prices of the instrument from the previous instrument history. For more details of nearest neighbour prediction capabilities refer to [18].

### 2.6 Expectation – Maximization Method

Expectation – maximization algorithm is a statistic iterative method used to find statistical model of the maximum likelihood estimates [19]. It is a suitable tool when experiment has no completeness of the training data for example missing values in the river records data set [6]. This method estimates an unobservable population parameter and maximizes the log-likelihood function (see Equation 5).

$$L(\theta, X) = \sum_{i=1}^N \log p(x_i | \theta) \tag{5}$$

Observations  $X = \{x_i | i=1, \dots, N\}$  are independent variables. The algorithm is an iterative process and with each iteration solves parameters estimation problem. Method finds missing values  $Y = \{y_i | i=1, \dots, n\}$  corresponding to  $X$  variables.

We assume that parameter  $\theta_0$  is known. Each algorithm iteration has two steps: expectation (E) step and the maximization (M) step.

1. In the expectation (E) step, the algorithm determines the expectation of log-likelihood the complete data based on the incomplete data  $Q(\theta | \theta^{(t)}) = E(\log p(X, Y | \theta) | X, \theta^{(t)})$
2. In the maximization (M) step, the algorithm determines a new parameter maximizing the equation  $\theta^{(t+1)} = \arg \max_{\theta} Q(\theta | \theta^{(t)})$

Each iteration guarantee to increase the likelihood estimates of the statistical model.

### 3 Experiments

There are several ways to construct artificial history. One way is to forecast the actual price. The other way is to forecast price changes and cumulatively sum them up. We analysed both methods. The price forecasting proved to be not very accurate. Therefore we chose to work with percentage price changes.

#### 3.1 The Data

We downloaded most liquid ETF series as well as NASDAQ and NYSE stocks (see in the Table 1) for the experiments.

**Table 1.** Source data information

Number of stocks used by methods	5
Total number of stocks	1000
Number of ETF for prediction	100
Dataset starts	1962-01-02
Dataset ends	2011-08-09

#### 3.2 The Flow of the Experiments

Creating accurate historical data is not an easy task. If created, how one could verify its validity? We selected 100 ETF's for the experiment. Also we selected 1000 shares as our source for ETF history reconstruction. For each ETF we selected 5 the most correlated stocks. It is based on our theoretical assumptions in section 2, Equation 1. We used the same stocks for all methods - all the methods had the same input data.

For performance evaluation we chose two different measures Mean Absolute Error (MAE) and Square Root of Mean Square Error (RMSE).

We also used in-sample and out-of-sample testing. For in-sample we used the most recent half (50 %) of the data and we predicted the older half (50 %). After constructing the model on the most recent half we applied it on the older half. Later, we compared predicted history with the actual history and computed MAE and RMSE.

To make our results more statistically significant, we repeated experiment with 100 different ETF series and averaged the results. Numbers in the result table are average of 100 experiments. We present standard deviation in brackets. One may use those statistics measures to compute statistical significance.

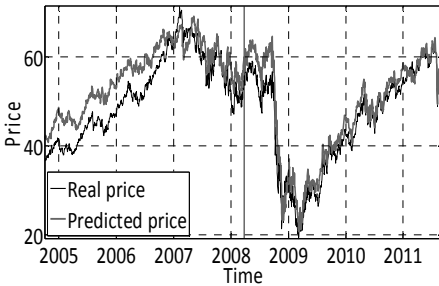
### 4 Results and Discussion

Main results of these experiments are presented in Table 2. We highlighted the winning model in bold which is expectation maximisation algorithm.

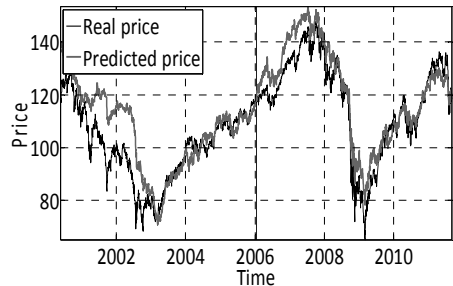
**Table 2.** Prediction accuracy in percent. Numbers in the table indicate that forecast on average differs from the actual value and is indicated in percent.

Method	In Sample (%)		Out of Sample (%)	
	RMSE (STD)	MAE	RMSE (STD)	MAE
Wiener Process	2.717 (1.4)	2.020	2.993 (2.0)	2.208
Proxy method	2.487 (1.1)	1.644	2.943 (2.0)	1.928
Multiple proxy method	2.451 (1.1)	1.642	2.895 (1.8)	1.949
Regression method with one stock history	1.912 (0.9)	1.176	2.379 (1.7)	1.429
Regression method with some stock histories (extended zeros)	2.158 (1.1)	1.466	2.431 (1.7)	1.487
Regression method with some stock histories (periodically predicted)	2.158 (1.1)	1.466	2.531 (1.8)	1.619
k - Nearest Neighbour method	2.200 (1.0)	1.524	2.539 (1.8)	1.707
<b>Expectation – Maximization Algorithm</b>	<b>1.838 (0.9)</b>	<b>1.106</b>	<b>2.223 (1.6)</b>	<b>1.212</b>
<i>Average</i>	<i>2.240 (1.0)</i>	<i>1.506</i>	<i>2.617 (1.8)</i>	<i>1.692</i>

For illustration purposes we present two successful cases (Fig. 4, 5) and two less successful cases (Fig. 6, 7) of history reconstruction using regression method.



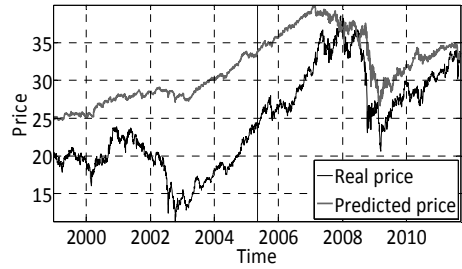
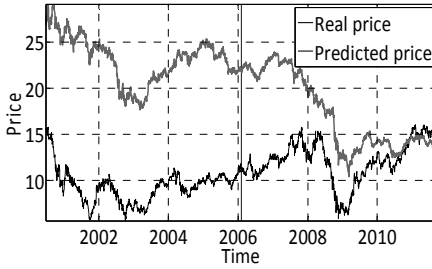
**Fig. 4.** ETF named VNQ history (2004-10-01 – 2011-09-08) prediction applied by regression method with a different model. Vertical line represents out-of-sample and in-sample boundary.



**Fig. 5.** ETF named IVV history (2000-05-19 – 2011-09-08) prediction applied by regression method with a different model. Vertical line represents out-of-sample and in-sample boundary.

Some ETF follows underlying index. Indexes are quite often composed from prices of some stocks, grouped with a specific weight. The problem is that composition of the index changes periodically. So stock weighting in the index can differ. Hence it is impossible to create precise linear model to produce index from stocks. The solution is approximate. The aim of this methodology is to be able to reconstruct historical prices approximately but very quickly, without the effort of analysing its components.

To solve the problem we suggest to extend the past ETF history using the most correlated stocks bucket. Despite various ETF history extension methods, all methods have capability to extend the history till the maximum as the longest stock history.



**Fig. 6.** ETF named EWT history (2000-06-23 – 2011-09-08) prediction applied by regression method with a different model. Vertical line represents out-of-sample and in-sample prediction.

**Fig. 7.** ETF named XLU history (1998-12-22 – 2011-09-08) prediction applied by regression method with a different model. Vertical line represents out-of-sample and in-sample prediction.

Unsuccessful cases in Fig. 6, 7 are examples where lack of correlated sources makes it difficult or impossible to reasonably accurately reconstruct missing history.

#### 4.1 Analysis

We present one synthetic history example in more detail. VNQ (“Vanguard REIT”) ETF is quoted in NYSE with a limited history that starts from 1<sup>st</sup> of October, 2004. Selected the most correlated stock data is shown in Fig 4. The whole history consists of 1748 days. Our methodology extends the history till the maximum length of the alternatives. In our example we select 5 stocks from 1000 stock basket. By selecting alternatives we assume that extension can be applied for 2200 days more. The longest stock history has 2200 days more than the VNQ itself.

**Table 3.** Similar stock basket used for history extension for VNQ

Stock ticker	Dataset start	History length (days)	Standard deviation	Correlation coefficient	Extension part (days)
BMR (“BioMed Realty Trust Inc.”)	2004-08-09	1786	4.22	0.7936	38
ACNB (“ACNB Corporation”)	1996-01-02	3948	2.49	0.7535	2200
AMZN (“Amazon.com”)	1997-05-16	3595	46.98	0.7527	1847
BDT (“BlackRock Strategic Dividend Achievers Trust”)	2004-05-06	1850	1.32	0.7460	102
BKYF (“Bank of Kentucky Financial Corp.”)	1999-02-25	3151	3.35	-0.7289	1403

We select the stock basket, that is the most similar to calculated ETF past history. Table 3 has more information about the sources.

We don't take stocks with a shorter history than our ETF. BMR and BDT have a very short history, so their influence on the results is very little. The longer ones will impact the prediction more. Note that AMZN standard deviation is so high comparing with others, therefore its impact would be different.

We use the same source data for the experiments with each prediction method. Methods use newer half of the history for the learning and the older half for verification of the results. Accuracy results are shown in Table 4.

**Table 4.** The VNQ history extension accuracy

Method	In-sample (%)		Out-of-sample (%)	
	RMSE (STD)	MAE	RMSE (STD)	MAE
Wiener Process	0.030 (0.3)	0.022	0.021 (0.2)	0.015
Proxy method	0.038 (0.4)	0.023	0.022 (0.2)	0.013
Multiple proxy method	0.035 (0.4)	0.023	0.025 (0.3)	0.014
<b>Regression method with a one stock history</b>	<b>0.024 (0.3)</b>	<b>0.012</b>	<b>0.010 (0.1)</b>	<b>0.006</b>
Regression method with some stock histories (extended zeros)	0.032 (0.4)	0.021	0.012 (0.1)	0.008
Regression method with some stock histories (periodically predict)	0.032 (0.4)	0.021	0.015 (0.2)	0.010
$k$ - Nearest Neighbour method	0.032 (0.3)	0.021	0.013 (0.1)	0.010
<b>Expectation - Maximization Algorithm</b>	<b>0.026 (0.3)</b>	<b>0.016</b>	<b>0.010 (0.1)</b>	<b>0.006</b>

In this example we see that the best prediction accuracy is achieved with a regression method with one stock history – using only one ACNB stock. This method is the second in the overall accuracy experiment see Table 2 (Out-of-sample, MAE). Another successful method is Expectation – maximization.

The worst result is obtained using a Wiener process. It is because the source data has no influence to the results. The method generates history by drawing it from normally distributed random set with the same mean and standard deviation as the ETF history. We also note that this method is the worst method overall in all experiments.

## 5 Conclusions

To increase the prediction accuracy in the future we examined the possibility of creating a synthetic past for exchange traded funds that have a limited amount of history. The synthetic history approach is rather new and has not been very widely discussed in the academic community. Furthermore, it is completely unknown in ETF trading. Historical data extension is particularly important for systematic trading funds that require a long history to test their automated trading systems.

To demonstrate the usefulness of the ETF time-series history reconstruction we used stocks with a longer history as source data. In the empirical study we compared eight



different methods: the Wiener process (random data), proxy, multiple proxy, regression with full history, regression with history of baskets, regression with zeros,  $k$ -nearest neighbour estimator and expectation maximization algorithm. For the first time we employed machine learning algorithms for ETF synthetic history creation.

Finally, we proved that a *synthetic history can be successfully created* and that the expectation maximization algorithm is the most suitable for this task. We showed both theoretically and experimentally that the gain in prediction accuracy relies on additional information contained in the data, which, however, is not used in the standard data mining methods: the asset similarity and correlation.

For the future, we plan to analyse other criteria for the selection of the source basket and new history reconstruction methods. The worst results produced methods will be replaced with more complex and hopefully better methods.

Our methodology can be used on any time series. For example, we can reconstruct a history of some stocks with a short trading history by using other stocks or any other sources like indexes. This, however, needs to be verified in our future works.

We also plan to make simulated investments using our generated data and see if this approach could add value in practice. Additionally we will investigate how reconstruction accuracy depends on the missing history length. We will also investigate a possible solution to minimize the percentage of unsuccessful cases.

**Acknowledgments.** This research was funded in part by a grant (No. MIP-018/2012) from the Research Council of Lithuania. The authors also want to express their appreciation to Vilnius University.

## References

1. Mockus, J., Raudys, A.: On the Efficient-Market Hypothesis and Stock Exchange Game Model. *Expert Systems with Applications* 37(8), 5673–5681 (2010)
2. Graham, K.: Imputing for missing survey responses. s.l. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association (1982)
3. Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* 14, 853–871 (2001)
4. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, pp. 3–18, 39–48, 127–139. John Wiley & Sons, Los Angeles (1987)
5. Tseng, S., Wang, K., Lee, C.: A pre-processing method to deal with missing values by integrating clustering and regression techniques. *Applied Artificial Intelligence* 17(5-6), 535–544 (2003)
6. Firat, M., Dikbas, F., Koc, A.C., Güngör, M.: Estimation of Missing River Flows using Expectation Maximization Method. Balwois, Ohrid (2010)
7. Amato, A., Calabrese, M., Di Lecce, V.: Decision Trees in Time Series Reconstruction Problems. In: *IEEE International Instrumentation and Measurement Technology Conference*, pp. 895–899. IEEE, Canada (2008)
8. Kim, J.-W., Pachepsky, Y.A.: Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT stream flow simulation. *Journal of Hydrology* 394(3-4), 305–314 (2010)

9. Huang, X., Zhu, Q.: A pseudo nearest neighbour approach for missing data recovery on Gaussian data sets. *Pattern Recognition Letters* 23(13), 1613–1622 (2002)
10. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Missing, Mining, Inference, and Prediction*. Springer, New York (2001)
11. Aksoy, H., Toprak, Z.F., Aytekin, A.: Stochastic generation of hourly mean wind speed data. *Renewable Energy* 29(14), 2111–2131 (2004)
12. Srikanthan, R.: A multisite daily rainfall data generation model for climate change conditions. In: 18th World IMACS / MODSIM Congress, pp. 3976–3982. eWater CRC, Water Division, Bureau of Meteorology, Melbourne (2009)
13. Andridge Rebecca, R., Little Roderick, J.A.: A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review* 78(1), 40–64 (2010)
14. Utsunomiya, K., Sonoda, K.: *Methodology for Handling Missing Values In Tankan*. Research and Statistics Department Bank of Japan, Japan (2001)
15. Bang, Y.-K., Lee, C.-H.: Fuzzy Time Series Prediction with Data Preprocessing and Error Compensation Based on Correlation Analysis. In: Third International Conference on Convergence and Hybrid Information Technology, vol. 2, pp. 714–721. IEEE (2008)
16. Shrestha, S.L.: Categorical Regression Models with Optimal Scaling for Predicting Indoor Air Pollution Concentrations inside Kitchens in Nepalese Households. *Nepal Journal of Science and Technology* 10, 205–211 (2009)
17. Sujatha, K.V., Sundaram, S.M.: Stock Index Prediction Using Regression and Neural Network Models under Non Normal Conditions. In: 2010 International Conference on Emerging Trends in Robotics and Communication Technologies (INTERACT), pp. 59–63 (2010)
18. Yankov, D., DeCoste, D., Keogh, E.: Ensembles of Nearest Neighbor Forecasts. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) *ECML 2006. LNCS (LNAI)*, vol. 4212, pp. 545–556. Springer, Heidelberg (2006)
19. Mustapha, N., Jalali, M., Bozorgniya, A., Jalali, M.: Navigation Patterns Mining Approach based on Expectation Maximization Algorithm. *World Academy of Science, Engineering and Technology* 50, 855–859 (2009)

# Dynamic Adaptive Algorithm Selection: Profit Maximization for Online Trading

Iftikhar Ahmad<sup>1,\*</sup>, Javeria Iqbal<sup>1</sup>, and Günter Schmidt<sup>1,2</sup>

<sup>1</sup> Saarland University, P.O. Box 151150, D-66041 Saarbrücken, Germany  
`{ia,ji}@itm.uni-sb.de`

<sup>2</sup> Department of Statistical Sciences, University of Cape Town, South Africa  
`gs@itm.uni-sb.de`

**Abstract.** Online trading algorithms can facilitate the investment decisions in financial markets. This paper presents a dynamic adaptive algorithm selection framework for online trading with the goal of maximizing overall revenue (lower competitive ratio). We integrate the algorithm selection (II) and probabilistic graphs to transform the algorithm selection model of offline algorithms for a dynamically adaptive model of online trading. Unlike the traditional static approach of algorithm selection, where a single algorithm is executed over the whole investment horizon, we dynamically select and update the trading algorithm by analyzing the time series features. The time series is partitioned in different windows and each window's features are extracted sequentially using real time hybrid pattern matching approach. The extracted features of current window  $w_i$  are analyzed and the decision making module determines the best suitable algorithm for next window  $w_{i+1}$  on the basis of underlying probabilistic graphical model. The process is repeated until all windows in a time series are processed. Our dynamic adaptive algorithm selection model outperforms the static model on real world datasets of *DAX30* and *S&P500*.

**Keywords:** Business applications of artificial intelligence, Business Intelligence, Online Trading, Adaptive Algorithms.

## 1 Introduction

Online algorithms can be used to facilitate the investments decisions in financial market such as foreign exchange. Online algorithms for conversion problems (e.g; foreign exchange) are designed for online scenarios where the input is revealed to the player in a sequential manner such that the player at any time  $t$  has no knowledge about the input at time  $t + 1$ . The decision of algorithm is based on the input sequence seen so far and the decision itself is irrevocable as well [2]. Competitive ratio is used as a performance measure to evaluate the online algorithms. It measures the performance of an online algorithm against an optimum offline algorithm. Let  $ON$  be an online algorithm for some maximization

---

\* Corresponding author.

problem  $\mathcal{P}$  and  $\mathcal{I}$  is set of all inputs. Let  $ON(I)$  be the return of algorithm  $ON$  on input instance  $I \in \mathcal{I}$ . Let  $OPT$  be the optimum offline algorithm for same problem  $\mathcal{P}$ , and  $OPT(I)$  its return for the input instance  $I \in \mathcal{I}$ . An online algorithm  $ON$  is called  $c$ -competitive if  $\forall I \in \mathcal{I}$ :

$$ON(I) \geq \frac{1}{c} \cdot OPT(I). \quad (1)$$

The traditional approach of applying algorithms for investment decisions pre-selects an algorithm which is executed over the whole length of the investment horizon. However, the approach may not be clever option as online algorithms are designed based on the competitive analysis, thus the performance of each algorithm is highly dependent on the input instance. As, an algorithm performing better on one set of input instances can perform really bad on another set of input instances. Thus in the absence of a “universally best” algorithm, selecting a single algorithm for the whole time series, solely on the basis of competitive ratio is not viable option. We modify the algorithm selection model [1] for online conversion problem in order to dynamically adapt the best algorithm over investment horizon based on the properties and behavior of input instance.

## 1.1 Problem Settings

### *Given*

1. A time series  $Q$ .
2. A set of online trading algorithms  $\mathcal{A}$ .
3. A set of patterns  $F$ .

### *Problem*

How to dynamically allocate/adapt the best trading algorithm based on the observed features in time series with the objective to maximize over all performance. A sub-problem in the domain is extraction of time series features  $f(Q) \in F$ .

### *Our Contribution*

We modify the Rice model [1] of (offline/static) algorithm selection for online conversion problems. Second, we extend the feature extraction approach of Zhang et al. [3] to accommodate identification of more patterns in time series. Third, the probabilistic graphical model is applied to identify patterns in time series for updating the pre-selected algorithm dynamically. In addition, we perform experiments on two real world datasets,  $DAX30$  and  $S\&P500$  to show the performance improvement using proposed dynamic approach against the traditional static approach.

## 2 Literature Review

The algorithm selection model selects the best performing algorithm per instance base as different algorithms perform well on different problem instances [1]. The

problem is discussed comparatively and quantitatively in the literature and applied to a number of optimization problems. Rice [1] investigated and discussed the applicability of approximation theory to the algorithm selection problem. Potkonjak and Rabaey [4] discussed the algorithm selection problem using quantitative optimization approach. They emphasized on optimization of throughput and cost using algorithm selection approach in system level design. Lagoudakis and Littman [5] applied the algorithm selection approach to dynamically choose algorithm per instance base with the objective to minimize the execution time. The authors formulated the problem as Markov decision process and used it in conjunction with reinforcement learning. Other related work includes Houstis et al. [6], Gagliolo et al. [7] and Xu et al. [8].

### 3 Proposed Framework

Our proposed framework is based on *Rice model* [1] of algorithm selection. The basic idea of Rice model is to select the most suitable algorithm based on problem instance's features. We present the theoretical formulation of algorithm selection problem by [1], followed by discussion on how we extend this model to propose a *Dynamic Adaptive Algorithm Selection Framework for Online Conversion Problems*.

The Rice model has 5 major modules as following:

1. **Problem Space**  $P$ , which constitutes of problem instances. A problem instance  $x$  is chosen from  $P$ ,  $x \in P$ .
2. **Feature Space**  $F$ , which contains features for each problem instance  $x \in P$ . A function  $f \in F$  is applied for feature extraction of an underlying problem instance  $x$ . The features  $f(x) \in F$  for pre-chosen problem instance  $x$  are utilized for determining the most appropriate algorithm. The decision about the most appropriate algorithm is based on the criteria space  $R^n$ .
3. **Criteria Space**  $R^n$ , which consists of performance rating parameters  $m \in R^n$ . The performance rating parameters depend upon user's choice, e.g., accuracy, efficiency, consistency. The selection mapping function  $S$  maps the extracted features and user's preferred performance rating parameters to an algorithm  $a \in A$ . The mapping function  $S$  is as follows:

$$S : F \times R^n \rightarrow A$$

4. **Algorithm Space**  $A$ , which contains a set of algorithms  $A$  for executing over pre-chosen problem instance  $x$ .
5. **Performance Space**  $p$ , which records the performance measures for each executed algorithm  $a \in A$  over problem instance  $x$ . The performance of an algorithm depends only on problem instance  $x$ . The performance mapping function is stated as follows:

$$p : A \times P \rightarrow R^n$$

The algorithm selection finds the best algorithm for a given problem instance according to the defined criteria space and populates the feature space. For example, Eald et al. [9] present Eq. 2 which finds the selection mapping  $S^*$  for nominating the best algorithm using extracted features  $f \in F$ , for a particular problem instance.

$$\forall x \in P, m \in R^n, a \in A$$

$$\|p(S^*(f(x), m), x)\| \geq \|p(a, x)\| \tag{2}$$

Fig. 1 highlights major modules of *Rice model*. Rice assumes that the input space is known to the model, which is a realistic assumption, as the model addresses offline NP-hard problems where complete problem instance is available before the algorithm starts execution, whereas our problem instance is not available before hand, and is revealed as the time progresses. Hence, our decision making regarding the most appropriate algorithm selection necessitates an online process. We modify the Rice model to address this issue. We discuss basic work flow of our proposed design and focus on its core building blocks, as presented in Fig. 2.

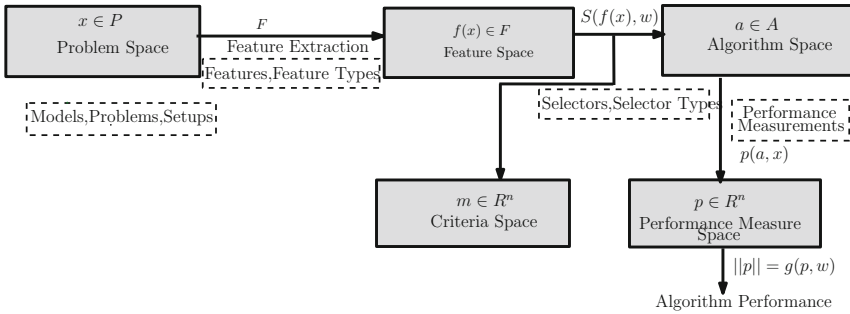


Fig. 1. Rice model of dynamic algorithm selection

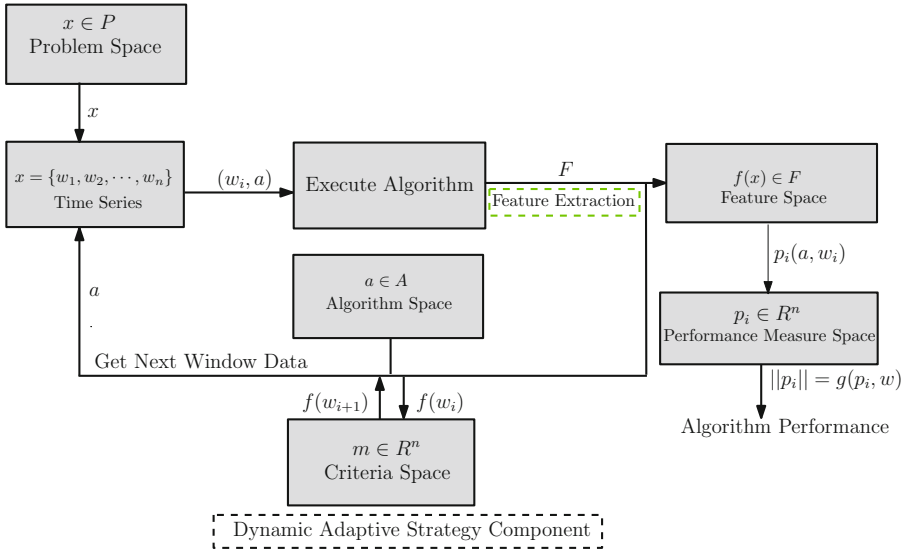
### Basic Work Flow

Our problem space  $P$  contains *time series' instances*. A time series instance from  $P$  is chosen and partitioned into equal sized windows of length  $l$  (last window's size may vary). The algorithm space  $A$  consists of *online trading algorithms*. An online trading algorithm  $a \in A$  is executed on first window and its features  $f(w_i)$  are extracted. The extracted features  $f(w_i)$  are forwarded to the criteria space. The criteria space defines a mapping between features  $f \in F$  and the best performing online trading algorithm  $a \in A$ . It identifies the features  $f(w_{i+1})$  for next window and selects the best algorithm for executing over next window  $w_{i+1}$ . This process is repeated for each window and online trading algorithms are updated according to the observed features.

### Core Building Blocks

The presented framework has two core components:

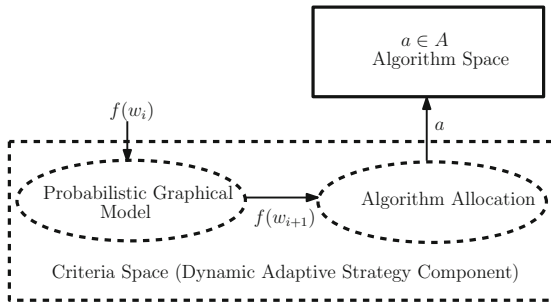
**i. Hybrid Feature Extraction Component:** The feature extraction involves extraction of interested patterns in time series. The component corresponds to



**Fig. 2.** Proposed model for dynamic adaptive algorithm selection

“Feature Space” of Rice model [1]. We apply the hybrid approach for finding patterns in time series data [3]. The approach consists of two phases. In first phase, we use *PIP-VD* to find perceptually important points and apply *Spearman rank correlation coefficient* for desired patterns’ classification. The Spearman rank correlation coefficient ranks the data, e.g., records the position of a data point in an ascending ordered list rather than the actual data values. This is a distribution independent scheme which maintains the integrity of input distributions and results in identification of desired patterns for a given time series.

However, to qualify the formation of a particular pattern, it must satisfy some rules. Hence, the second phase applies a set of pre-defined rules to further enhance the results and to eliminate patterns that do not match the specified



**Fig. 3.** Components of DASC

set of rules [3]. We extend the rule set to incorporate more patterns that are not considered by Zhang et al. [3]. The following set of patterns are considered.

- i.* Head & Shoulder
- ii.* Reverse-Head & Shoulder
- iii.* Double Top
- iv.* Reverse-Double Top
- v.* Spike Top
- vi.* Reverse-Spike Top
- vii.* Triple Top
- viii.* Reverse-Triple Top
- ix.* Rounded Top
- x.* Reverse-Rounded Top
- xi.* Up Trend
- xii.* Down Trend
- xiii.* Flag up
- xiv.* Flag Down

**ii. Dynamic Adaptive Strategy Component:** The Dynamic Adaptive Strategy Component (DASC) dynamically adapts the best algorithm after a specified time interval. The decision of algorithm selection is based on two sub-components, namely next pattern identification via “probabilistic graphical model” and algorithm selection via “algorithm allocation” as shown in Fig. 3.

**a. Next Pattern Identification:**

Next pattern identification is based on the pre-constructed directed probabilistic graphical model (PGM) [10]. Due to space constraint, we only define the probabilistic graphical model and briefly discuss its construction process.

Factor is a basic unit of PGM. Each factor is a boolean valued function over input random variables. Hence, the PGM captures uncertainty and correlations using these boolean valued functions. Let  $R$  denotes a random variable and  $domain(R)$  denotes the domain of corresponding random variable.

**Definition 1.** A factor  $f(R)$  is a function over a set of random variables (base facts)  $R = R_1, \dots, R_n$  where  $f(r) \in \{0, 1\}, \forall r \in domain(R_1) \times \dots \times domain(R_n)$ . We denote a set of random variables with  $S(R)$ .

**Definition 2.** A probabilistic graphical model (PGM)  $P = (F, S(R))$  computes a joint distribution over a set  $S(R)$  with corresponding factors  $F_{factor}^i$ , where  $\forall f(R) \in F_{factor}, R \subseteq S(R)$ . Given a complete joint assignment  $r \in domain(R_1) \times domain(R_2) \dots domain(R_n)$  to variables in  $S(R)$ . The joint distribution is defined by Eq. 3.

$$Pr(r) = \frac{1}{Z} \prod f(r_f) \tag{3}$$

Where  $r_f$  denotes assignments restricted to arguments of  $f$  and  $Z$  is a normalization constant represented by Eq. 4.

$$Z = \sum_{r'} \prod_{f \in F} f(r'_f) \tag{4}$$

The PGM is constructed from  $S$ , where  $S$  is a set of times series from historical data,  $S = \{Q_1, Q_2, \dots, Q_k\}$ . Each node in probabilistic graphical model is a random variable which represents an underlying pattern (feature)  $f_i$ . A sample PGM is shown in Fig. 4 which demonstrates the dependencies among different random variables. The probability of each node in PGM is based on Markov



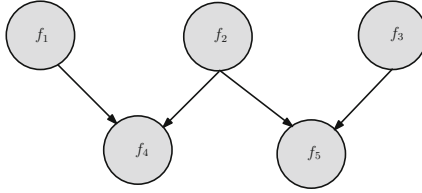


Fig. 4. A sample PGM

condition. The Markov condition for our directed *PGM* enforces that a node is conditionally independent of its non-predecessors, e.g., node  $f_4$  in Fig. 4 is conditionally independent of each other node except  $f_1$  and  $f_2$ . The probability of occurrence for feature  $f_4$  can be calculated by the following formula:

$$P(f_4 | \text{Predecessors of } f_4) = P(f_4 | f_1, f_2)$$

Using the Baye’s rule:

$$P(f_4 | f_1, f_2) = \frac{P(f_1, f_2 | f_4) \cdot P(f_4)}{P(f_1, f_2)}$$

The criteria space is referred as *Dynamic Adaptive Strategy Component (DASC)*, as shown in Fig. 3. The Hybrid Feature Extraction Component extracts features  $f(w_i)$  for each window  $w_i$ . The extracted features of each window  $f(w_i)$  are passed to the first sub-component of *DASC* named as *PGM*. It identifies features for next window  $f(w_{i+1})$  that can occur in  $w_{i+1}$  based on an appeared pattern  $f(w_i)$ . The identified features are forwarded to the second sub-component called as algorithm allocation.

**b. Algorithm Allocation**

The algorithm allocation depends on the criteria space. The criteria space is populated before framework starts its execution. We create synthetic dataset for each pattern and execute each algorithm  $a \in A$  on synthetic datasets. The performance of each algorithm  $a \in A$  is observed for each pattern  $f \in F$ . A mapping between each occurred feature  $f$  and the best performing algorithm  $a$  is stored in the algorithm allocation, sub-component of criteria space. The algorithm allocation component receives the identified pattern for next window  $f(w_{i+1})$  by *PGM*. It finds the mapping function  $S^*$  for next window  $w_{i+1}$  based on the identified patterns  $f(w_{i+1})$ . This mapping function finds the best performing algorithm for  $w_{i+1}$  based on the identified patterns  $f(w_{i+1})$ , as shown by Eq. 5.

$$\forall x \in P(\forall w_i \in x, m \in R^n, a \in A),$$

$$||p_i(S^*(f(w_i), m), w_i)|| \geq ||p_i(a, w_i)|| \tag{5}$$

Where  $i = \{1, 2, 3, \dots, n\}$  and  $n$  denotes the number of windows. The mapping functions are identified and the best performing algorithm is selected for the next window  $w_{i+1}$ . The process is repeated until all windows are processed.

## 4 Experimental Results

This section presents experimental settings and methodology followed by evaluation of results. The performance in terms of competitive ratio improves significantly using proposed framework with dynamic adaption capability over static approach for algorithm selection which increases the overall performance.

### 4.1 Settings

We use two real world datasets for performance verification of our proposed framework: *DAX30 (2001 to 2010)* and *S&P500 (2001 to 2010)*. We partition the algorithm space  $A$  in two categories: *Progressive Algorithms* and *Non-Progressive Algorithms*.

**Progressive Algorithms:** The algorithms which do not convert at each offered price but only convert when offered price is the highest seen so far. We discuss algorithms suggested by El-Yaniv et al. [2] and Lorenz et al. [11]

**Threat based Algorithm by El-Yaniv et al. [2]**

*Given  $M, m$  or  $M, m, k$ :* El-Yaniv et al. [2] discuss four variants of Algorithm [1]. Each variant assumes different knowledge about future. We consider two variants of threat based strategy of El-Yaniv et al. [2] (YFKTMm and YFKTMmk).

Where,  $m$  and  $M$  are the lower and upper bounds for offered prices, and  $k$  denotes the day after which the offered price may drop to some minimum level  $m$ . For the experiments, we substitute  $k = T$ .

**Algorithm 1.** *Three basic rules of threat-based algorithm are as follows:*

1. *Consider a conversion from asset  $D$  into asset  $Y$  only if offered price is the highest seen so far.*
2. *Whenever convert asset  $D$  into asset  $Y$ , convert just enough  $D$  to ensure that a competitive ratio  $c$  is obtained if an adversary drops offered price to minimum possible price  $m$ , and minimum price  $m$  is offered later on.*
3. *On last trading day  $T$ , all remaining  $D$  must be converted into  $Y$ , possibly at price  $m$ .*

**$u$ -Preemptive Algorithm by Lorenz et al. [11]**

*Given  $M, m$ :* Lorenz et al. [11] present a  $u$ -preemptive strategy with known  $m$  and  $M$ . Two strategies are proposed: A selling strategy known as *Max-search Problem* and a buying strategy known as *Min-search Problem*. We discuss the max-search (selling) algorithm.

**Algorithm 2.** *Max-search (selling) Problem:*

*At the start of game reservation prices  $q_i^* = (q_1^*, q_2^*, \dots, q_u^*)$ , where  $i = 1, \dots, u$  are*

computed. The adversary unfolds prices and algorithm accepts the first price which is at least  $q_1^*$ . The player waits for next price which is at least  $q_2^*$ , and continues in the same way. If some units of asset are left on last day  $T$ , then all remaining units must be sold at the last offered price, even if it is the lowest price  $m$ .

$$q_i^* = m \left[ 1 + (c^* - 1) \left( 1 + \frac{c^*}{u} \right)^{i-1} \right] \tag{6}$$

Here,  $c^*$  represents competitive ratio for max-search (selling) problem.

**Theorem 1.** Let  $u \in N$ ,  $\phi > 1$ , there exists a  $c^*$ -competitive deterministic algorithm for  $u$  max-search problem where  $c^* = c^*(u, \phi)$  is the unique solution of

$$\frac{(\phi - 1)}{(c^* - 1)} = \left( 1 + \frac{c^*}{u} \right)^u. \tag{7}$$

**Non-progressive Algorithms:** The algorithms which do not consider any trends or local maxima criterion for conversion and convert a portion of initial wealth at each offered price. We consider Chen et al. [12] and Hu et al. [13] algorithms.

**Non-progressive Algorithm by Chen et al. [12]**

**Given**  $T, g(q_t)$ : Chen et al. [12] present an algorithm for uni-directional search. The algorithm assumes prior knowledge of duration  $T$ , and price function  $g(q_t)$ . The constants  $A$  and  $B$  ( $A, B \geq 1$ ) determine the prices offered on a day  $t$ , and  $q_t$  is modeled as  $q_{t-1}/B \leq q_t \leq A \cdot q_{t-1}$ .

The algorithm and amount invested  $s_t$  on day  $t$  is described as follows:

**Algorithm 3.** Determine amount to convert at time  $t$  by the following rules:

$$s_t = \begin{cases} \frac{A(B-1)}{TAB - (T-1)(A+B) + (T-2)} & t = 1 \\ \frac{(A-1)(B-1)}{TAB - (T-1)(A+B) + (T-2)} & t \in [2, T-1] \\ \frac{(A-1)B}{TAB - (T-1)(A+B) + (T-2)} & t = T. \end{cases} \tag{8}$$

**Theorem 2.** The competitive ratio  $c$  achieved by Algorithm 3 is

$$c = \frac{TAB - (T-1)(A+B) + (T-2)}{AB - 1} \tag{9}$$

**Non-progressive Algorithm by Hu et al. [13]**

**Given**  $g(q_t), T$ : Hu et al. [13] presents two algorithms which assume prior knowledge of price function  $g(q_t)$  and duration of investment horizon  $T$ . First algorithm called as *static mixed strategy* is overly pessimistic. It assumes worst case input of price sequences and fixes competitive ratio which does not change. Thus, a second algorithm called as *dynamic mixed strategy* is proposed which considers number of remaining days  $T' = T - t + 1$ , and price function  $g(q_t)$  for conversion. Competitive ratio is improved by recalculation of achievable competitive ratio. The price function is modeled as  $(1 - \gamma)q_{t-1} \leq q_t \leq (1 + \gamma)q_{t-1}$ .

**Static Mixed Strategy:** The static mixed strategy allocates amount to be converted based on worst-case input sequence of prices.

**Algorithm 4.** Determine amount to convert at time  $t$  by the following rules:

$$s_t = \begin{cases} \left( \frac{1+\gamma}{(T-1)\gamma+2} \right) & t = 1 \\ \left( \frac{\gamma}{(T-1)\gamma+2} \right) & t \in [2, T - 1] \\ \left( \frac{1}{(T-1)\gamma+2} \right) & t = T \end{cases} \quad (10)$$

**Theorem 3.** The competitive ratio  $c$  achieved by Algorithm 4 is

$$c = 1 + \frac{\gamma}{2} (T - 1) \quad (11)$$

**Dynamic Mixed Strategy:** The dynamic mixed strategy allocates  $s_t$  based on remaining number of days  $T'$  in time interval  $T$ .

**Algorithm 5.** Determine amount to convert at time  $t$  by the following rules:

$$s_t = \begin{cases} \left( \frac{1+\gamma}{(T'-1)\gamma+2} \right) W'_t & t = 1 \\ \left( \frac{\gamma}{(T'-1)\gamma+2} \right) W'_t & t \in [2, T - 1] \\ \left( \frac{1}{(T'-1)\gamma+2} \right) W'_t & t = T \end{cases} \quad (12)$$

where  $W'_t$  denotes remaining amount of wealth at day  $t$ .

**Theorem 4.** The competitive ratio  $c$  achieved by Algorithm 5 based on remaining number of days  $T'$  is

$$c = 1 + \frac{(T' - 1)\gamma}{2}. \quad (13)$$

## 4.2 Methodology

The experiments are performed on yearly data. Competitive ratio is used as performance measure and is calculated for each algorithm on the given data. We calculate the average competitive ratio of each algorithm for both datasets: DAX30 and S&P500. For the sake of simplicity, we consider the transaction cost and interest rate to be zero. All prices considered are closing day prices. First we execute the individual algorithms on the data sets (DAX30 and S&P500) and for each year record the average competitive ratio achieved by the set of algorithms. This is followed by execution of our proposed dynamic algorithm selection model on given datasets. The experimental evaluation methodology is presented in Algorithm 6.

**Algorithm 6.** Experimental Methodology

**Input:** Dataset  $D$  with  $y$  number of years, window length  $l$ .

1. **for each** ( $y \in D$ ) **do**
2. Divide each year  $y$  into  $n$  windows of equal length  $l$ , last window's length may vary.
3. Randomly choose an online trading algorithm 'a' from algorithm space  $A$ .
4.  $i \leftarrow 1$
5. **while** ( $i \leq n$ ) **do**
  - i. Execute algorithm 'a' on window  $w_i$
  - ii. Extract features  $f$  from window  $w_i$  using "Hybrid Feature Extraction Component"
  - iii. Forward extracted features  $f$  to Dynamic Adaptive Strategy Component (DASC) as an input.
  - iv. Identify next pattern by DASC using probabilistic graphical model and return  $f(w_{i+1})$ .
  - v. Select the best performing algorithm  $a \in A$  based on  $f(w_{i+1})$  for next window  $w_{i+1}$ .
  - vi.  $i \leftarrow i + 1$
6. **End while**
7. **End for**

### 4.3 Results and Discussion

Fig. 5 summarizes the results for DAX30 and S&P500 datasets. It is important to note that the objective is to achieve a lower competitive ratio(cf. Eq. 1). For each year, the average competitive ratio of algorithm space (static algorithm selection) and that of dynamic algorithm selection is listed. In each year, the dynamic algorithm selection outperforms the static algorithm selection approach. On dataset DAX30, the performance of dynamic algorithm selection is 9% better than static algorithm selection, with minimum performance difference of 5.84%, in year 2007 and maximum performance difference of 12.36%, in year 2005.

We observe the same performance edge using dynamic algorithm selection for S&P500 dataset as for DAX30 dataset. The dynamic algorithm selection consistently out-performs the static algorithm allocation. The average performance difference on S&P500 is 7%. In addition, we consider the best performing algorithm from algorithm space for each year on both DAX30 and S&P500 datasets, and compare its performance with dynamic algorithm selection approach. The results reveal that performance of our approach is 4% better on average than the corresponding best performing algorithm from algorithm space.

In order to show the consistency of our proposed approach, we use variance of the competitive ratio. Variance can be used to show the consistency of the applied approach, the lower the variance, the more consistent the approach is. Table 1 summarizes the variance of static and dynamic approach on DAX30 as well as S&P500 data sets. It can be deduced from Table 1 dynamic approach is more consistent as it has considerably lower variance on both data sets.

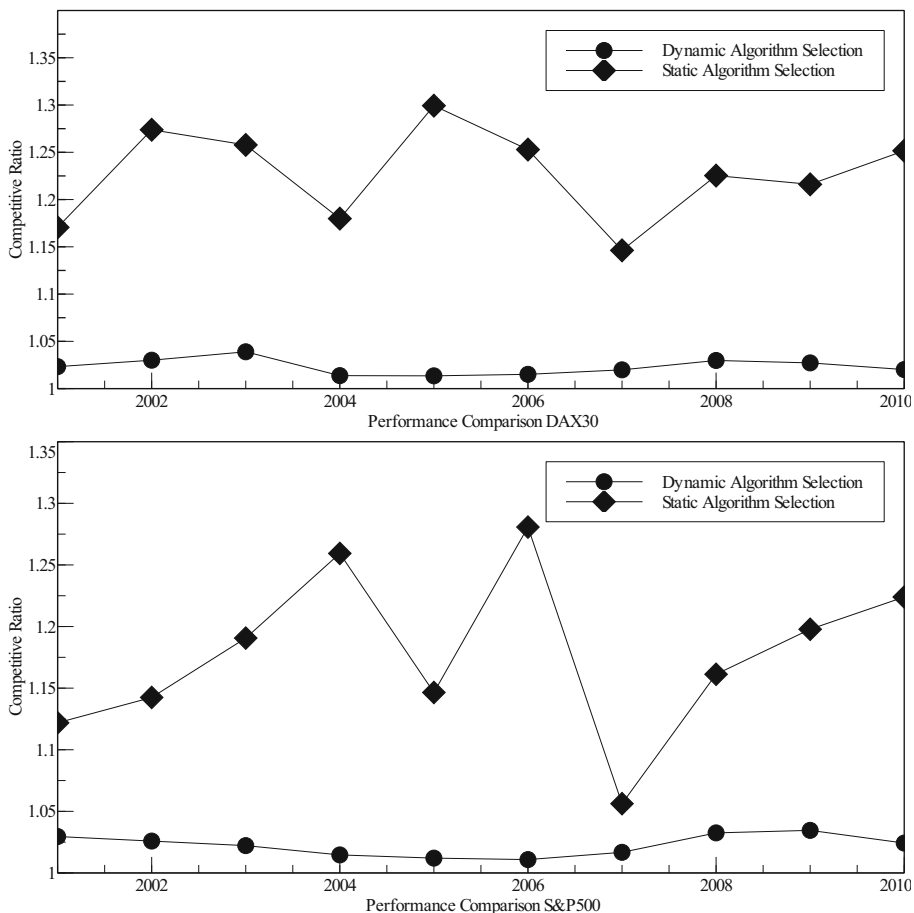


Fig. 5. Performance comparison on DAX30 and S&P500 datasets: static vs. dynamic approach

Table 1. Variance: static vs. dynamic approach

Data Set	Static Approach ( $10^{-3}$ )	Dynamic Approach ( $10^{-3}$ )
DAX30	2.4102	0.0696
S&P500	4.4837	0.0722

## 5 Conclusion

The traditional algorithm selection approach is static. It pre-selects an online trading algorithm and executes it over whole input instance without considering the properties of input instance. We present a profit maximization framework for algorithm selection in online conversion problems. The presented design is a

modified form of Rice model [11]. Possible future direction includes investigating the proposed model for bi-directional conversion problems, where a player is allowed to convert an asset back and forth to maximize the overall return after time  $T$ . A comparative study of inference and learning models can provide useful insight for profit maximization and decision making in online trading environment, e.g., statistical and probabilistic models with genetic programming and neural networks.

## References

1. Rice, J.R.: The algorithm selection problem. *Advances in Computers* 15, 65–118 (1976)
2. El-Yaniv, R., Fiat, A., Karp, R.M., Turpin, G.: Optimal search and one-way trading algorithm. *Algorithmica* 30, 101–139 (2001)
3. Zhang, Z., Jiang, J., Liu, X., Lau, R., Wang, H., Zhang, R.: A real time hybrid pattern matching scheme for stock time series. In: *Proceedings of the Twenty-First Australasian Conference on Database Technologies, ADC 2010*, vol. 104, pp. 161–170. Australian Computer Society, Inc., Darlinghurst (2010)
4. Potkonjak, M., Rabaey, J.: Algorithm selection: a quantitative computation-intensive optimization approach. In: *ICCAD 1994: Proceedings of the 1994 IEEE/ACM International Conference on Computer-Aided Design*, pp. 90–95. IEEE Computer Society Press, Los Alamitos (1994)
5. Lagoudakis, M.G., Littman, M.L.: Algorithm selection using reinforcement learning. In: *Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000*, pp. 511–518. Morgan Kaufmann Publishers Inc., San Francisco (2000)
6. Hazan, E., Seshadhri, C.: Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity (ECCC)* 14(088) (2007)
7. Gagliolo, M., Zhumatiy, V., Schmidhuber, J.: Adaptive online time allocation to search algorithms, pp. 134–143 (2004)
8. Xu, L., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Satzilla: portfolio-based algorithm selection for sat. *J. Artif. Int. Res.* 32, 565–606 (2008)
9. Ewald, R., Himmelspach, J., Uhrmacher, A.M.: An algorithm selection approach for simulation systems. In: *Proceedings of the 22nd Workshop on Principles of Advanced and Distributed Simulation, PADS 2008*, pp. 91–98. IEEE Computer Society, Washington, DC (2008)
10. Jordan, M.I.: Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)* 19, 140–155 (2004)
11. Lorenz, J., Panagiotou, K., Steger, A.: Optimal algorithms for  $k$ -search with application in option pricing. *Algorithmica* 55(2), 311–328 (2009)
12. Chen, G.-H., Kao, M.-Y., Lyuu, Y.-D., Wong, H.-K.: Optimal buy-and-hold strategies for financial markets with bounded daily returns. *SIAM Journal on Computing* 31(2), 447–459 (2001)
13. Hu, S., Guo, Q., Li, H.: Competitive Analysis of On-line Securities Investment. In: Megiddo, N., Xu, Y., Zhu, B. (eds.) *AAIM 2005*. LNCS, vol. 3521, pp. 224–232. Springer, Heidelberg (2005)

# Credit Risk Modeling of USA Manufacturing Companies Using Linear SVM and Sliding Window Testing Approach

Paulius Danenas and Gintautas Garsva

Department of Informatics, Kaunas Faculty, Vilnius University, Muitines St. 8,  
LT- 44280 Kaunas, Lithuania  
{paulius.danenas,gintautas.garsva}@khf.vu.lt

**Abstract.** This paper presents a study on credit risk evaluation modeling using linear Support Vector Machines (SVM) classifiers, combined with feature selection and “sliding window” testing approach. Discriminant analysis based evaluator was applied for dynamic evaluation and formation of bankruptcy classes. The research demonstrates a possibility to develop and apply an intelligent classifier based on original discriminant analysis method evaluation and shows that it might perform bankruptcy identification even better than original model.

**Keywords:** Support Vector Machines, linear SVM, machine learning, credit risk, evaluation, bankruptcy.

## 1 Introduction and Related Research

Company classification by their risk can be described as one of the key components of credit risk evaluation model. It plays an important role in decision process of acceptance/rejection projects of credit application. This problem applies through analysis of various financial and other customer data to conclude the final decision. Sophisticated and effective tools to solve task must be developed. The combination of machine learning and statistical techniques might help to minimize the drawbacks of separate techniques and thus develop models which might prove more accurate than common statistical techniques. This research proposes a technique which is based on popular machine learning technique, namely Support Vector Machines. The proposed method is also tested in “sliding window” approach manner, which means that it can be useful to identify more general trends. Moreover, the combination of this method with discriminant analysis (or other similar techniques) might be useful while trying to improve the performance of these methods by identifying the most relevant financial attributes and developing a new classifier based on that particular technique.

The application of intelligent classification techniques in credit risk domain is dated back to 1968 when Altman et. al. [1] applied discriminant analysis. They obtained 96% and 79% accuracy by using two different samples, however, it is



reliable in its predictive ability only in two years, after that the results fall down significantly. Zmijewski [2] applied probit (simple probit and bivariate) and maximum likelihood principles to a set of 40 bankrupt and 800 non-bankrupt companies and a prediction sample of 41 bankrupt and 800 non-bankrupt companies collected from American and New York Stock Exchanges, resulting in 72% accuracy for complete dataset case. Springate [3] developed his model using step-wise multiple discriminate analysis to select 4 ratios which best describe a failing company. It obtained an accuracy rate of 92.5% using the 40 companies tested by Springate; later 83.3% and 88% accuracy rates were reported after testing it with other samples [4]. Ohlson used logit approach to construct his model [5], and he reported accuracy of 96.12%, 95.55% and 92.84% for prediction within one year, two years and one or two years respectively.

Support Vector Machines is applied for efficient classification obtaining results comparable to Neural Networks and other machine learning techniques. As for credit risk domain, they were successfully applied for company failure prediction [6], financial warning prediction [7], to evaluate financial fate of Dotcoms [8], rating companies [9], to estimate probability of default [10], to study credit rating systems [11], capital risk assessment [12]. Lai and Zhou proposed several SVM based methods for various credit risk related tasks, such as identification of high-risk customers [13] or credit scoring [14]. These authors also developed several Least Squares SVM (LS-SVM) based methods, including their developed Weighted LS-SVM techniques [15] [16] or LS-SVM ensemble models [16][17]. LS-SVM integration into credit risk process was also researched by van Gestel et al. in their works [18][19]; they showed that LS-SVM can provide better performance in both complexity and accuracy. These authors also combined it with Bayesian evidence framework for regularization and kernel parameter selection to predict financial distress of Belgian and Dutch firms with middle market capitalization [20].

A model for forecasting changes which combines discriminant analysis technique together with a supervised neural network applied to increase performance in terms of accuracy has been proposed in [21]. This model was applied to forecast changes in discriminant models although it may be applied to forecast changes in ratings or expert evaluations as well.

SVM has been intensively researched in this field with combination with various soft computing techniques; the advantages and disadvantages of these combinations are described in [22]. Danenas and Garsva [23] tried to combine SVM classification technique with discriminant analysis for credit risk evaluation. Their research showed that LIBLINEAR and SMO algorithms are capable to obtain results similar to Vapnik's SVM classifier results. A comparative research of various SVM classifiers by these authors [24] proved that linear SVM classifiers can be a good alternative for credit risk evaluation model development in terms of both complexity and speed, in case there is no need for nonlinear separation using complex kernel functions.

## 2 The Method

### 2.1 Description of Algorithms Used in this Experiment

**Support Vector Machines.** Support Vector Machines is an efficient and effective solution for pattern recognition problem whereas a following minimization problem has to be solved in order to generate weight vector:

$$\min - \sum_{i=1}^{\ell} \alpha_i + \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to } \sum_{i=1}^{\ell} y_i \alpha_i = 0, \forall i: 0 \leq \alpha_i \leq C$$

where the number of training examples is denoted by  $l$ , training vectors  $X_i \in R, i = 1, \dots, l$  and a vector  $y \in R^l$  such as  $y_i \in [-1; 1]$ .  $\alpha$  is a vector of  $l$  values where each component  $\alpha_i$  corresponds to a training example  $(x_i, y_i)$ . If training vectors  $x_i$  are not linearly separable, they are mapped into a higher (maybe infinite) dimensional space by the kernel function  $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$  where classifier is generated by minimizing an appropriate convex cost function. This can be done in Support Vector Machines (SVM), Least Squares SVM (LS-SVM) [18][19] and other kernel based learning techniques, such as kernel regression or kernel PCA (which is used for extraction of linearly uncorrelated variables). Then the solution is obtained in the dual space from a finite dimensional convex quadratic programming problem for SVM or a linear Karush–Kuhn–Tucker system in the case of LS-SVM, avoiding explicit knowledge of the high dimensional mapping and using only the related positive (semi) definite kernel function [20].

**LIBLINEAR.** LIBLINEAR is an open source library and a family of linear SVM classifiers for large-scale linear classification which can be very efficient for training large-scale problems. These classification methods do not use kernel functions for transformation into other space which makes it possible train a much larger set much faster.

Formally these algorithms (except Crammer and Singer algorithm) are defined as follows: given training vectors  $x_i \in R^n, i = 1, \dots, l$  in two class, and a vector  $y \in R^l$  such that  $y_i = \{1, -1\}$ , a linear classifier generates a weight vector  $w$  as the model using a decision function

$$\text{sgn}(w^T x)$$

One-vs-All (OVA) strategy is used for multiclass classification problems; that is, for  $K$ -class problem,  $K$  binary classifiers are built separating one class from the rest, and the answer is chosen according to the hyperplane which separates the point with the highest confidence from other data points.

An approach proposed by Crammer and Singer for solving an optimization problem is based on multiclass classification, thus it is defined differently [25]: given training vectors  $x_i \in R^n, i = 1, \dots, l$  and a vector  $y \in R^l$  such that  $y_i \in \{1, \dots, k\}$  a weight vector is generated using

$$\arg \max_{m=1,\dots,k} w_m^T x$$

Table 1 gives formulations of these algorithms (algorithms and primary optimization problems that are solved); more information is given in [25].

**Table 1.** Definitions of the algorithms used in research

Algorithm	Minimization problem
L2-regularized L1-loss SVC	$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^l (\max(0, 1 - y_i w^T x_i))$
L2-regularized L2-loss SVC	$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^l (\max(0, 1 - y_i w^T x_i))^2$
L2-regularized logistic regression	$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^l \log(1 + e^{-y_i w^T x_i})$
L1-regularized L2-loss SVC	$\min_w \ w\ _1 + C \sum_{i=1}^l (\max(0, 1 - y_i w^T x_i))^2$  ( $\ \cdot\ _1$ defines 1-norm)
Multi-class SVM by Crammer and Singer	$\min_{w_m, \xi_i} \frac{1}{2} \sum_{m=1}^k w_m^T w_m + C \sum_{i=1}^l \xi_i$  subject to $w_{y_i}^T x_i - w_m^T x_i \geq e_i^m - \xi_i, i = 1, \dots, l$  $e_i^m = 0, \text{ if } y_i = m$ $e_i^m = 1, \text{ if } y_i \neq m$

These classifiers also include a bias term  $b$ , which handled by augmenting the weight vector  $w$  and each instance  $x_i$  with an additional dimension. An interesting and useful notice is that all these classifiers have a considerably low number of additional parameters (i.e., only cost parameter  $C$  and bias parameter  $b$ ) which makes it easier to choose appropriate classifier parameters.

## 2.2 Research Methodology

This research applies modified method proposed in [23][24], using classifiers defined in previous section.. The modified algorithm is defined as follows:

1. Evaluate each financial entry by using discriminant analysis (or any other expert evaluation method, if possible) and compute bankruptcy classes.
2. Eliminate instances which could not be evaluated in Step 1 because of lack of data or division by zero and thus resulted in empty outputs.
3. Remove attributes from the dataset which have less values than specified threshold (70% was considered in this case).
4. Data imputation is performed by filling missing values with average value of particular attribute.
5. Perform the following steps for each  $m \in [1, n - k]$ , where  $n$  is the total number of periods,  $k$  is the number of periods which are used for forecasting:
  - a. Apply feature selection procedure in order to select the most relevant attributes and reduce number of dataset dimensions;
  - b. Perform classifier parameter selection manually or using heuristic procedures;
  - c. Train classifier using data from first  $m$  periods.
  - d. Apply hold-out testing using data from period  $p$ ,  $p \in [m + 1, m + k]$ ;  $p \in \mathbb{N}$ .

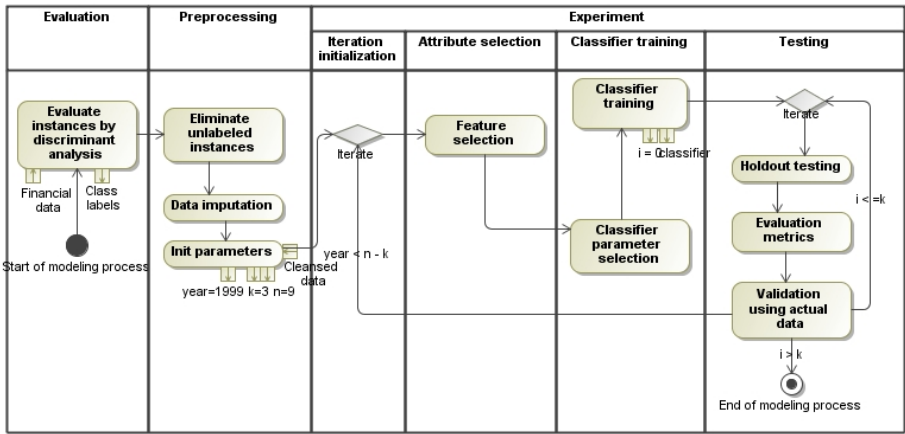
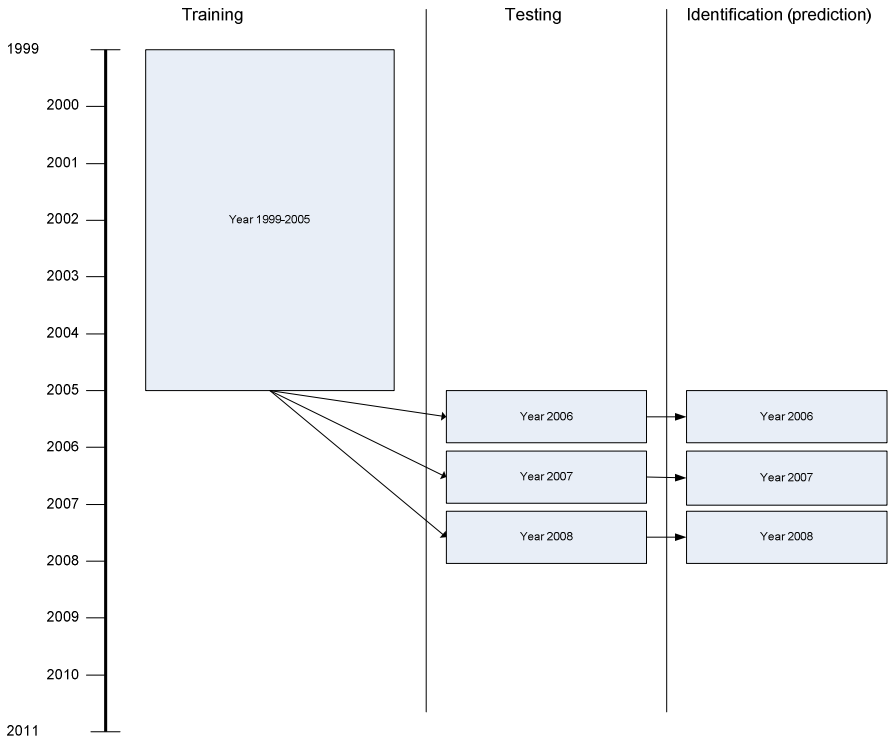


Fig. 1. Workflow of method used in experiment

Figure 1 represents the algorithm graphically as a workflow. The output (for each iteration in experimental stage) is the trained classifier (list of support vectors in case of SVM) and the list of selected attributes.

Finally, to test model performance, an additional step is performed using real bankruptcy data. If applied dataset is in the period  $[p_{start}; p_{end}]$ , with year  $p_{end}$  as the year of last entry in financial history, bankruptcy is known to be occurred after the financial history, i.e., on year  $p_{end} + 1, p_{end} + 2, \dots, p_{end} + k_y$ , with  $k_y$  as the maximum number of years during which the company is officially recognized as bankrupt.



**Fig. 2.** Sliding window approach for testing and bankruptcy identification

Figure 2 gives a graphical overview of overall method, which comprises training, testing and identification (prediction) stages. In this picture,  $k = 3$  and  $k_y = 3$  (this combination is used in experiment). I.e., the data from 1999 to 2005 is used for training, the developed model is used for testing data from year 2006, year 2007 and year 2008 individually). After testing with year 2006, bankruptcy identification is performed on this year: the instance in financial history record representing year  $p_{end}$  is labeled as “Risky” (as it was bankrupt), and prediction procedure is performed on the instance.

### 3 The Experiment

#### 3.1 Research Data

The experiments were made by using data from EDGAR database, manufacturing sector, from year 1999-2008. The initial dataset used in the experiment consists of yearly financial records with 51 financial ratios used in financial analysis; these ratios were computed using original primary financial data from balance and income statement data.

**Table 2.** Main characteristics of data used in experiment

Year	Entries labeled as		Total entries	No of selected attributes	Bankrupt 1 years after	Bankrupt >1 year after
	Risky (R)	Not risky (NR)				
1999	1312	537	1849	12	-	-
2000	1869	589	2458	15	0	0
2001	1753	672	2425	15	1	0
2002	1709	777	2486	13	3	0
2003	1770	723	2493	14	0	2
2004	1920	637	2557	13	0	1
2005	1964	660	2624	14	3	17
2006	1636	429	2065	14	0	3
2007	1545	393	1938	14	1	13
2008	483	109	592	14	4	1
<b>Total</b>	<b>15961</b>	<b>5527</b>	<b>21487</b>		<b>12</b>	<b>37</b>

Table 2 presents main characteristics of dataset, including classes formed by evaluation using Zmijewski's score, together with bankruptcy data from UCLA database used to validate the results. UCLA LoPucki database [30] contains bankruptcy data and covers about 50 companies from used dataset. The data from 2000 – 2010 period was applied for validation; instances which represent last entry in financial history were marked as “risky” and were evaluated by developed classifiers.

### 3.2 Experiment Configuration

Zmijewski's score [2] was used in this research as an evaluator to form class labels and formulate the problem as a classification problem; it was selected because of the origin of the data (which comes from USA and Canada companies). This scoring technique allows to form two groups of companies – companies which are “healthy” (possibly are not going to bankrupt) and “bad” (which might become bankrupt). Zmijewski's score is defined as follows:

$$Z = -4,336 - 4,513 * (\text{Net Revenue} / \text{Total Assets}) + 5,679 * (\text{Total Debt} / \text{Total Assets}) + 0,004 * (\text{Short Term Assets} / \text{Short Term Assets})$$

If  $Z < 0$  then company is considered as “risky” (prone to bankruptcy).

The code and algorithms for the experiments was implemented using Weka framework [28] with LIBLINEAR 1.7. The cost parameter  $C$  and bias  $b$  were chosen experimentally, by using grid search in range of  $C \in [0;100]$  and  $b \in [0;1]$ . Feature selection was applied for each formed dataset using correlation-based feature subset evaluation [29] to select the most relevant financial ratios.

For the Step 6 of our procedure, it is presumed that bankruptcy might have occurred following the year of the last entry of financial history for particular company, next year or even later ( $k$  years after).  $k = 3$  is selected in this experiment; thus bankruptcy fact is evaluated here only if it happens during the next 3 years after the last entry in financial records of the company.

The results were evaluated using accuracy, True Positive Rate (TPR) and F-Measure. These metrics are often used in machine learning and more information about them can be found in various sources, such as [27] where these metrics were used to evaluate results.

### 3.3 Experiment Results

Table 3 presents the classification results - classifier parameters, classification accuracy together with TPR and F-Measure rates for each class. The accuracy is above 90%, which can be considered as very good result. Best results were obtained using different classifiers – Cramer-Singer multiclass SVM showed best performance for 2 analyzed cases, L1 dual linear SVM – for 4 cases and L2 linear SVM, both primal and dual – for last two cases (once per each classifier). Thus different classifiers obtained best results for different periods.

**Table 3.** Results of experiment

Training period		2000	2001	2002	2003	2004	2005	2006	2007	
Structure (parameters)		CS-SVM	L1-LSVM (dual)	L1-LSVM (dual)	CS-SVM	L1-LSVM (dual)	L1-LSVM (dual)	L2-LSVM (primal)	L2-LSVM (dual)	
C		20	20	20	15	20	15	15	5	
Bias		0.7	1.0	0.7	1.0	0.4	0.7	0.7	1.0	
Year 1	Accuracy	<b>96,702</b>	<b>96,344</b>	<b>95,471</b>	<b>95,504</b>	<b>91,604</b>	<b>93,085</b>	<b>92,008</b>	<b>92,295</b>	
	TP	R	0,973	0,974	0,970	0,965	0,974	0,977	0,971	0,981
		NR	0,951	0,940	0,917	0,925	0,745	0,756	0,724	0,675
	FMeas	R	0,977	0,973	0,968	0,970	0,945	0,957	0,951	0,954
		NR	0,941	0,942	0,922	0,911	0,818	0,820	0,789	0,770
Year 2	Accuracy	<b>96,183</b>	<b>94,233</b>	<b>95,348</b>	<b>96,785</b>	<b>92,940</b>	<b>91,445</b>	<b>91,960</b>	-	
	TP	R	0,966	0,966	0,972	0,983	0,977	0,966	0,977	-
		NR	0,953	0,938	0,898	0,923	0,749	0,716	0,675	-
	FMeas	R	0,972	0,970	0,969	0,979	0,956	0,947	0,952	-
		NR	0,940	0,928	0,906	0,936	0,816	0,775	0,762	-
Year 3	Accuracy	<b>96,032</b>	<b>96,286</b>	<b>96,710</b>	<b>97,389</b>	<b>91,291</b>	<b>92,127</b>	-	-	
	TP	R	0,962	0,970	0,987	0,987	0,964	0,981	-	-
		NR	0,956	0,940	0,908	0,923	0,716	0,667	-	-
	FMeas	R	0,972	0,975	0,978	0,984	0,946	0,953	-	-
		NR	0,933	0,927	0,933	0,936	0,772	0,764	-	-

The TPR values for both “risky” (R) and “non-risky” (NR) classes were high (both were over 0.9 in first four periods, and over 0.7 in next periods); this shows that instances for both of these classes were recognized separated and procedures for unbalanced learning were not needed to apply. High F-Measure values which are more suitable for unbalanced learning evaluation also prove this. Parameters C and

bias varied; the experiment showed that bias parameter had significant influence thus further research targeted at parameter selection might show even better results.

Table 3 shows that best classification results were obtained while training classifier sequentially with data from first five years (starting with year 1999). Classification resulted in accuracy over 95%. Later it decreased, although the number of instances used for training increased. This might indicate a trend of changes in the data, as well as overall financial situation change; yet the classification performance remained above 90%.

**Table 4.** Bankruptcy prediction results

Year	Number of actual bankrupt	Original model (Zmijewski)	No of bankruptcies after testing period		
			Year 1	Year 2	Year 3
2002	1	0	0	-	-
2003	3	0	0	0	-
2005	2	0	0	0	0
2006	4	1	1	1	1
2007	1	0	1	0	0
2008	8	6	6	5	5
2009	27	9	18	16	17
2010	3	0	1	1	1
<b>Total:</b>	<b>49</b>	<b>16</b>	<b>27</b>	<b>23</b>	<b>24</b>

The last step was performed in order to compare the performance of proposed approach with the performance of the original model. Table 4 represents identification (prediction) results. As this table shows, the model developed by the proposed method identified more bankruptcy facts than original Zmijewski model which was used as the evaluator. This might mean that additional statistically selected predictors improved the performance and identified more bankruptcies than the original model in which ratios were selected on the basis of their performance in prior studies. The results varied for each year; yet overall performance was better. Note that Table 2 shows there were far more financial ratios considered relevant by feature selection procedure than the ones that were used in original evaluator. This proves that usage of higher dimensional data and more complex model might result in improved results.

## 4 Conclusions and Further Research

An approach for credit risk evaluation using linear SVM classifiers, combined with feature selection and sliding window testing is presented in this article. The classifier used here is based on linear SVM classifiers which are perfectly suitable for large scale learning. The developed classifiers were applied for real-world dataset, together with widely applied Zmijewski technique as a basis for output formation. This approach could serve as a alternative tool for company classification in case when



there are no actual bankruptcy classes as well as if obtaining them might be a too complicated or expensive. The classifiers were rerun on datasets based on the same principle as described above. Model validation was performed on real bankruptcy list; the obtained results showed that it outperformed the original Zmijewski model.

One of the main problems related to proposed method is possible imbalanced learning arising from the fact that classes are computed dynamically by external evaluator. Although this research did not have to deal with this problem integration of such procedure would be a useful complementary step. This is crucially important in identification of hazardous companies if they are represented by minority entries, as identification of hazardous company might cost more to the creditor than the misidentification of it.

## References

1. Altman, E.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* 23(4), 589–609 (1968)
2. Zmijewski, M.: Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* 22, 59–82 (1984)
3. Springate, G.L.V.: Predicting the Possibility of Failure in a Canadian Firm. Unpublished M.B.A. Research Project, Simon Fraser University (1978)
4. Sands, E.G., Springate, G.L.V., Var, V.: Predicting Business Failures. *CGA Magazine*, 24–27 (May 1983)
5. Ohlson, J.A.: Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research* 18(1), 109 (1980)
6. Yang, Z.R.: Support Vector Machines for Company Failure Prediction. In: *Proceedings of 2003 IEEE International Conference on Computational Intelligence for Financial Engineering*, pp. 47–54 (2003)
7. Wang, X.: Corporate Financial Warning Model Based on PSO and SVM. In: *2nd International Conference on Information Engineering and Computer Science (ICIECS)*, pp. 1–5 (2010)
8. Bose, I., Pal, R.: Using Support Vector Machines to Evaluate Financial Fate of Dotcoms. In: *Proceedings on Pacific Asia Conference on Information Systems (PACIS)*, Paper 42 (2005)
9. Hardle, W.K., Moro, R., Schafer, D.: Rating Companies with Support Vector Machines. *DIW Berlin, Diskussionspapier*, 416 (2004)
10. Hardle, W., Moro, R., Schafer, D.: Estimating Probabilities of Default With Support Vector Machines. *Discussion Paper Series 2: Banking and Financial Studies*. Deutsche Bundesbank, Research Centre (2008)
11. Chen, W.-H., Shih, J.-Y.: A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications* 30, 427–435 (2006)
12. Chong, W., Yingjian, G., Dong, W.: Study on Capital Risk Assessment Model of Real Estate Enterprises Based on Support Vector Machines and Fuzzy Integral. In: *Control and Decision Conference*, pp. 2317–2320 (2008)
13. Lai, K.K., Yu, L., Wang, S.-Y., Huang, W.: An Intelligent CRM System for Identifying High-Risk Customers: An Ensemble Data Mining Approach. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) *ICCS 2007*. LNCS, vol. 4488, pp. 486–489. Springer, Heidelberg (2007)

14. Lai, K.K., Zhou, L., Yu, L.: A Two-Phase Model Based on SVM and Conjoint Analysis for Credit Scoring. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007, Part II. LNCS, vol. 4488, pp. 494–498. Springer, Heidelberg (2007)
15. Zhou, L., Lai, K.K.: Weighted LS-SVM Credit Scoring Models with AUC Maximization by Direct Search. In: Proceedings of 2009 International Joint Conference on Computational Sciences and Optimization, pp. 7–11 (2009)
16. Zhou, L., Lai, K.K.: Multi-Agent Ensemble Models Based on Weighted Least Square SVM for Credit Risk Assessment. In: Proceedings of 2009 WRI Global Congress on Intelligent Systems, pp. 559–563 (2009)
17. Zhou, L., Lai, K.K., Yu, L.: Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications* 37, 127–133 (2010)
18. Van Gestel, T., Baesens, B., Garcia, I.J.: A support vector machine approach to credit scoring (2003), [http://www.defaultrisk.com/pp\\_score\\_25.htm](http://www.defaultrisk.com/pp_score_25.htm)
19. Van Gestel, T., Baesens, B., Suykens, J.A.K., Espinoza, M., Baestaens, D.E., Vanthienen, J., De Moor, B.: Bankruptcy prediction with least squares support vector machine classifiers. In: Proceedings of the International Conference on Computational Intelligence for Financial Engineering, CIFER, pp. 1–8 (2003)
20. Van Gestel, T., Baesens, B., Suykens, J.A.K., Van Den Poel, D., Baestaens, D.E., Willekens, M.: Bayesian kernel based classification for financial distress detection. *European Journal of Operational Research* 172, 979–1003 (2006)
21. Merkevičius, E., Garšva, G., Simutis, R.: Neuro-discriminate Model for the Forecasting of Changes of Companies Financial Standings on the Basis of Self-organizing Maps. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4488, pp. 439–446. Springer, Heidelberg (2007)
22. Danenas, P., Garsva, G.: Support Vector Machines and their Application In Credit Risk Evaluation Process. *Transformations in Business & Economics* 8(3(18)), 46–58 (2009)
23. Danenas, P., Garsva, G.: Credit Risk Evaluation Using SVM-Based Classifier. In: Abramowicz, W., Tolksdorf, R., Węcel, K. (eds.) BIS 2010, Part 1. LNBIP, vol. 57, pp. 7–12. Springer, Heidelberg (2010)
24. Danenas, P., Garsva, G., Gudas, S.: Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. In: Proceedings of the International Conference on Computational Science (ICCS 2011), *Procedia Computer Science*, vol. 4, pp. 1699–1707 (2011)
25. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research* 9, 1871–1874 (2008)
26. Zmijewski, M.: Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* 22, 59–82 (1984)
27. Danenas, P., Garsva, G., Simutis, R.: Development of Discriminant Analysis and Majority-Voting Based Credit Risk Assessment Classifier. In: Proceedings of the 2011 International Conference on Artificial Intelligence, ICAI 2011, Las Vegas, Nevada, JAV, July 23–27, vol. 1, pp. 204–209 (2011)
28. Weka 3: Data Mining Software in Java, <http://www.cs.waikato.ac.nz/ml/weka/>
29. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning, Hamilton, New Zealand (1998)
30. UCLA-LoPucki Bankruptcy Research Database, <http://lopucki.law.ucla.edu/index.htm>

# A Context Framework for Process-Oriented Information Logistics\*

Bernd Michelberger<sup>1</sup>, Bela Mutschler<sup>1</sup>, and Manfred Reichert<sup>2</sup>

<sup>1</sup> University of Applied Sciences Ravensburg-Weingarten, Germany  
{bernd.michelberger,bela.mutschler}@hs-weingarten.de

<sup>2</sup> Institute of Databases and Information Systems, University of Ulm, Germany  
manfred.reichert@uni-ulm.de

**Abstract.** A continuously increasing data overload makes it a challenging task for knowledge-workers and decision-makers to quickly identify relevant information, i.e., information they need when executing business processes. To tackle this challenge, *process-oriented information logistics* is a promising approach. The basic idea is to provide the right process information, in the right format and quality, at the right place, at the right point in time, and to the right people. To achieve this, it becomes particularly important to take the work context of process participants into account. In fact, knowing and utilizing context information is a pre-requisite to effectively provide relevant process information to process participants. This paper provides a sophisticated *context framework* for enabling context-awareness in process-oriented information logistics.

**Keywords:** process-oriented information logistics, context-aware delivery of process information.

## 1 Introduction

Nowadays, enterprises are faced with a continuously increasing amount of data [1]. Knowledge-workers and decision-makers suffer from this data overload, since it makes it difficult for them to identify and access the needed information to perform their current tasks in the best possible way [2]. In the following, we call this information "process information", i.e., process information is information supporting process participants when working on business processes. Examples include e-mails, office files, best practices, or process descriptions. In practice, however, this alignment is difficult to accomplish since process information is typically handled separately from business processes and their execution [3].

To close this gap, *process-oriented information logistics* (POIL) is a promising approach. Goal is to provide the right process information, in the right format and quality, at the right place, at the right point in time, and to the right people. More precisely, POIL enables the process-driven, context-aware delivery of

---

\* This paper was done in the niPRO research project. The project is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 17102X10. More information can be found at <http://www.nipro-project.org>.

process information to knowledge-workers and decision-makers. POIL is particularly suitable for knowledge-intensive business processes involving large amounts of process information, user-interaction, and decision-making.

Various approaches have been proposed to enable POIL, including *data warehousing* and *business intelligence*. However, these approaches have not primarily been designed with POIL in mind. Data warehousing, for example, rather focuses on the creation of an integrated database [4]. Opposed to this, POIL focuses on the management of process information flows to support the execution of business processes. Traditional business intelligence, in turn, addresses data analytics and is typically completely isolated from business processes execution [3]. Moreover, information supply is often restricted to decision-makers. Conversely, POIL focuses on integration and analysis of process information as well as their delivery to both knowledge-workers and decision-makers.

What has been neglected so far is the support of knowledge-workers and decision-makers by providing personalized and contextualized process information. The latter is required to address the different needs of process participants. For example, less experienced process participants need more detailed information than experienced ones. To enable such differentiation, a process participant's context needs to be identified. For this purpose, his or her situation is described according to its characteristics, so-called *context information*. Besides process-related context information (e.g., process step, temporal process constraints), user-related context information (e.g., user name, role, experience level), device-related context information (e.g., display size, bandwidth), location-based context information (e.g., position), time-based context information (e.g., current date), and environment-related context information (e.g., temperature, noise level) may be considered as well. This paper proposes a context framework for POIL and the handling of context information to support the context-aware delivery of process information being relevant for process participants.

The presented research is performed in the niPRO project. In this project we apply semantic technology to integrate process information within *process information portals*. Our overall goal is to support knowledge-workers and decision-makers with the process information needed depending on their current working context. Key challenges include the provision of contextualized process information, flexible visualization of process information [5], and the development of design approaches for different levels of process information quality [2].

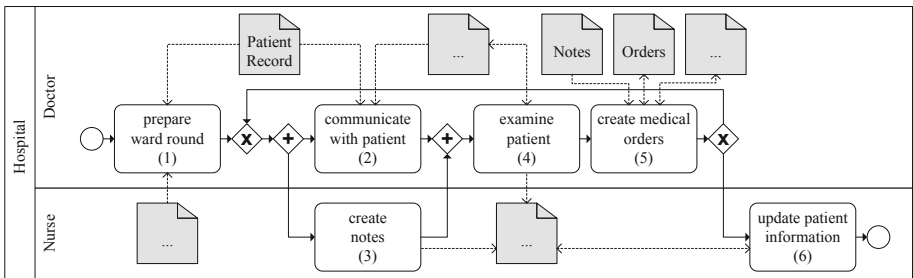
This paper is organized as follows. Section 2 gives a motivating example. Section 3 motivates the need for context-awareness in POIL and section 4 presents our context framework. Section 5 discusses related work. Finally, section 6 concludes the paper with a summary and an outlook.

## 2 Motivating Example

We use a scenario from the clinical domain, to motivate our approach. This scenario is based on lessons learned during an exploratory case study we performed at a German university hospital [6]. In this case study we analyzed the process

of an unplanned, stationary hospitalization, including patient admission, medical indication in the anesthesia, surgical intervention, post-surgery treatment, patient discharge, and financial accounting & management.

Our scenario (cf. Fig. 1) focusses on the ward round. First, the ward round is prepared (1), i.e., the doctor scans patient information and current medical instructions (e.g., endoscopic investigations, physical therapies). After finishing initial preparations, the doctor visits his patients. The doctor communicates with a patient and asks for information about his status (2). This information is written down by a nurse in parallel (3). Afterwards, the patient is examined (4). This activity includes the analysis of blood values and further follow-up diagnosis. Then the doctor creates medical orders (5). Finally, a nurse updates patient information and initiates further medical orders (6).



**Fig. 1.** Motivating example: Ward round

For each of the six process steps a variety of process information is needed. For example, to perform the process step "create medical orders" (5) a doctor needs access to blood values, notes, and current medical orders. Note that the mentioned process information only constitutes a small part of all processed information. In practice, there exist numerous different process information distributed across data sources (e.g., databases, shared drives, Intranet portals) [6]. Typical process information include, for example, process descriptions, working guidelines, operational instructions, forms, checklists, and best practices (e.g., documented in text documents, spreadsheets, and e-mails) [2].

As discussed, it is a big challenge to align process information with business processes. To reach this goal, different facets of POIL have to be addressed. In a first step, we have investigated issues related to process information quality [2]. In this paper, we take a closer look at context-awareness in POIL and propose a context framework.

### 3 Context-Aware Information Delivery

Context-awareness in POIL aims at the context-aware delivery of process information being relevant for a process participant. We adopt the notion of Dey [7]

and define *context-awareness* in POIL in a general way: POIL uses context information to deliver relevant process information to process participants, where relevancy depends on the participant’s task and process information quality requirements such as completeness or granularity [2].

Generally, context-awareness in POIL comprises three basic aspects: *sensors*, *context*, and *situation* (cf. Fig. 2). Reconsider our motivating example (cf. Fig. 1) and assume that a doctor performs process step “communicates with patient”. Thus, the *situation* (*S*) at hand could be described as follows: “Doctor Peter Miller communicates with a patient on Monday, 12th November, 2011, in room number 301 using a tablet computer”. Based on this we are able to characterize the situation using context information. For example, process-related context information (e.g., process step: “communicates with patient”), user-related context information (e.g., first name: “Peter”, last name: “Miller”, role: “doctor”), time-based context information (e.g., weekday: “Monday”, day: “12th”, month: “November”, year: “2011”), location-based information (e.g., room number: “301”), and device-related context information (e.g., used device: “tablet computer”) can be utilized.

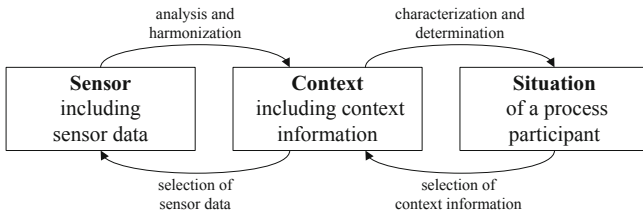


Fig. 2. Relationship between situation, context, and sensors

Particular context information (e.g., last name, role, day) is called *context aspect* (*CA*). Context aspects are further classified into different categories (e.g., process, user, time) denoted as *context factors* (*CF*). The set of all context factors is called *context* (*C*) (cf. Fig. 3). In order to determine specific values of context aspects, *sensors* (*SE*) are required. For example, to determine the context aspect “temperature” the sensor “thermometer” can be used. In the following we take a closer look at the three basic aspects (cf. Fig. 2).

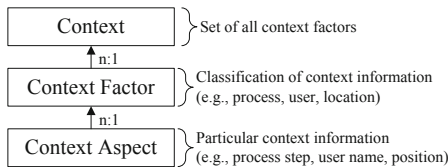


Fig. 3. Relationship between context, context factors, and context aspects

### 3.1 Sensors

The term *sensor* is specified by Haseloff as "any hardware or software systems that provides data about the entire or a part of the context of one or more entities" [8]. According to Indulska and Sutton [9] sensors can be classified into three categories: *physical sensors* (e.g., thermometer, microphone), *virtual sensors* (e.g., keyboard input, touch display movement, database trigger), and *logical sensors* (e.g., detect a process participant's position by analyzing logins at devices and a mapping of devices to locations). The main task of a sensor is to provide sensor data representing the initial value of the context aspects (e.g., the lightning sensor identifies the value of the context aspect lightning).

Based on this characterization, we can give a formal definition of the term *sensor*. Let  $SE$  be the sensor and  $v$  be the sensor value. We distinguish between *simple* (cf. Formula (1)) and *logical* (cf. Formula (2)) sensors. For example, a simple sensor can be a Global Position System (GPS) module determining the current position of a process participant. A logical sensor, in turn, can be a software system determining the user name based on first name and last name.

$$SE_{\text{simple}} := v \quad (1)$$

$$SE_{\text{logical}} := \{v_1, v_2, \dots, v_n\}, \quad n \geq 2 \quad (2)$$

### 3.2 Context

The notion of *context* as defined by Schilit et. al [10] or Pascoe et. al [11] is too specific and is based on an explicit set of context factors. Hence, these definitions become problematic when additional, so far unconsidered context factors need to be considered. In our research we need a more dynamic composition of context factors depending on the situation and on available sensors.

Depending on the process participant's situation different context factors are relevant. For example, to be able to update patient information (6) through the nurse it is not important to know where the task is performed. Conversely (e.g., in a case of an emergency) it is important to know which doctor is nearest to the emergency department. Therefore, we define context in a more general way according to Dey [7]: Context is any information that may be used to characterize the situation of an entity. The latter may be a person, location, or object being considered as relevant for the interaction between a process participant and a process information portal including the process participant and process information portal themselves.

We can now provide a formal definition of the terms *context*, *context factor*, and *context aspect*: Let  $C$  be the context,  $CF$  be the context factor,  $CA$  be the context aspect, and  $SE$  the sensor.  $C$ ,  $CF$ , and  $CA$  can be defined as follows:

$$C := CF_1 \cup CF_2 \cup \dots \cup CF_n \quad (3)$$

$$CF := \{CA_1, CA_2, \dots, CA_n\} \quad (4)$$

$$CA := \{SE_1, SE_2, \dots, SE_n\} \quad (5)$$

### 3.3 Situation

Finally, a *situation* can be characterized as "the world state at an instant of time" [12][13]. Haseloff more accurately says that "a situation is a part of the world state at a specific point in time or within a specific time interval" [8]. In other words, a situation represents the instantiation of the context at an instant of time. However, to describe the situation of a process participant we do not need the whole world state, but only those parts which might be relevant for POIL. Let  $S$  be the situation,  $C$  the context,  $t_{\text{start}}$  the starting time of the situation, and  $t_{\text{end}}$  its end time.  $S$  can then be defined as follows:

$$S := \langle C, t_{\text{start}}, t_{\text{end}} \rangle \quad (6)$$

The next section presents our context framework to enable context-aware delivery of process information.

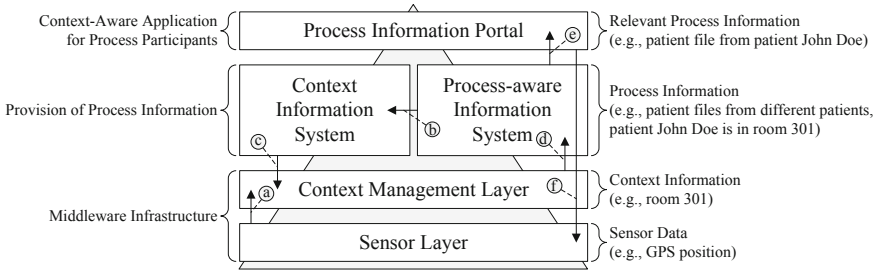
## 4 Context Framework

Our context framework aims at the context-aware delivery of relevant process information to process participants. It has been influenced by mobile and ubiquities computing and is therefore applicable to mobile scenarios as well (e.g., a ward round). The fundamental difference between our framework and existing ones is the explicit consideration of business processes. In fact, existing frameworks strongly focus on geographic services (e.g., provide the local temperature based on a current position) but do not address important ideas of POIL as discussed in Section 3. When compared to existing frameworks, our context framework does not directly provide any context information to applications (e.g., a weather application). Instead, it utilizes context information to determine the process information a process participant needs. More precisely, our context framework deals with gathering, representing, storing, analyzing, and providing of context information in order to enable the context-aware delivery of process information. As a consequence, existing frameworks can only be partially transferred. For the remainder of this paper we introduce the architecture of the framework and an ontology-based context modeling approach handling and representing context information in a machine-interpretable form.

### 4.1 Context-Framework Architecture

Generally, a context framework can be based on different architectures depending on business requirements. Chen [14], for example, distinguishes between three different architectural designs: *direct access to sensors*, *context server*, and *middleware infrastructure*. We adopt the latter viewpoint for several reasons, e.g., the reduced complexity resulting from the reduced number of data connections as well as the separation of business logic from the presentation layer and the data layer. Figure 4 illustrates the layered architecture of our context framework.





**Fig. 4.** Architecture of the context framework

The *sensor layer* is responsible for the management of sensor data (e.g., temperature, keyboard input) collected by different sensors (e.g., thermometer, keyboard). The sensor layer provides logical functionality, for example, functions to identify the role of a user by analyzing his or her access rights. Furthermore, the sensor layer allows for adding, removing, and switching (e.g., the GPS module will be replaced by a radio-frequency identification (RFID) system) sensors as well as encapsulating sensor communication (i.e., applications do not directly access sensor data).

The *context management layer* manages context information. Its main components include a context management layer interface, a context analytic engine, and a context model (not shown in Figure 4). The context management layer interface enables retrieval of sensor data from the sensor layer and provision of context information to higher layers via public interfaces. The context analytic engine allows for reasoning, interpreting, and aggregating context information (e.g., instead of GPS coordinates, the specific room number is given) [15]. Finally, the context model is responsible for storing and handling context information (cf. Section 4.2).

The *context information system* provides process information (e.g., which device belongs to which user, hospital map) to enrich available context information. The context information system can be seen as a support application. In the area of mobile computing, a geographic information system (GIS) has similar goals, but is limited to geographically information.

The *process-aware information system* (PAIS) contains integrated process information to support the execution of business processes, i.e., by delivery of process information to process participants. Its task is to gather process information from different data sources (e.g., databases, applications, shared drives), to analyze this process information (e.g., by using text similarity, usage pattern), and to offer it via public interfaces to other applications. Other functions include monitoring, event handling, and process information security.

The *process information portal* is responsible for the context-aware delivery of relevant process information to knowledge-workers and decision-makers.

Figure 4 also shows the dependencies between the different architectural layers. The sensor layer provides sensor data (e.g., user name, current process step) to the context management layer (a). Simultaneously, the context information

system provides certain process information (e.g., inventory lists, building maps) to the context management layer (b). The context information system obtains its process information from the PAIS (c). Based on the data/information flows of (a) and (b), the context management layer identifies context information and makes it available to the PAIS (d). The PAIS, in turn, uses this context information to identify relevant process information. The latter is then provided to the process information portal (e), which offers relevant process information to employees. Besides, the process information portal can be a sensor for the sensor layer (e.g., in order to gather user actions, clickstreams) (f).

The next section deals with the representation and handling of context information in ontology-based context model.

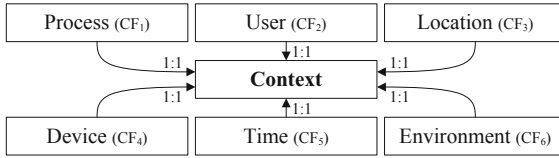
## 4.2 Context Model

The *context model* constitutes a fundamental part of our context framework. We use the context model to store and handle context information in a machine-interpretable form. Table 1 summarizes fundamental requirements (R1-R10) for such context models, which were gathered based on a literature study, two exploratory case studies [6], and an additional online survey [16].

**Table 1.** Requirements for a context model in POIL

#	Requirement
R1	The context model should represent all context information being relevant to the process participant's situation.
R2	The context model should be able to hide irrelevant context information in specific situations (e.g., location in non-mobile scenarios).
R3	The context model should be flexible and scalable to cope with the challenges of different update rates of context information.
R4	The context model should enable an efficient context analytic (e.g., reasoning, interpreting, aggregation) of context information.
R5	The context model should allow storing and handling historical context information.
R6	The context model should allow an efficient handling (e.g., fast processing, easy accessibility) of context information.
R7	The context model should be combined with the process information in order to provide contextualized process information
R8	The context model should be able to interpolate context information, to cope with incomplete context information.
R9	The context model should be easy to use, so that applications designers can easily translate real-world information to context information.
R10	The context model should store context information taking into account privacy and security issues.

Generally, a context model for POIL should represent all context information being relevant in the current situation of a process participant. However, a context model has to be restricted because the set of context information is infinite [17]. Indeed, any context modeling approach can only capture some parts of all possible context information. Hence, what we require is a classification of context information which allows reducing the complexity of context modeling. For instance, context information (e.g., first name, last name) in the same category (e.g., user) can be processed using the same or similar algorithms. We use the following six context factors (cf. Fig. 5) in POIL.



**Fig. 5.** Our Context Factors

- *Process* ( $CF_1$ ) includes process-related context aspects and reflects on what is currently (and in the past) happening. This includes, for example, general process aspects (e.g., process schemas, process instances, goals), time-based process aspects (e.g., durations and time lags between process steps, restricting execution times), responsibility-based process aspects (e.g., process owner), and data-based process aspects (e.g., input and output files).
- *User* ( $CF_2$ ) includes user-related context aspects and reflects who is involved in a certain situation. Thereby, we distinguish between explicit user aspects (e.g., user name, first name, last name, birthday, department) and implicit ones (e.g., experience, interests).
- *Location* ( $CF_3$ ) includes location-based context aspects and reflects where a situation takes place. This context factor includes both physical (e.g., GPS coordinates, Geolocation, and RFID systems) and logical (e.g., meeting room or office room) location aspects.
- *Device* ( $CF_4$ ) includes device-related aspects and reflects which devices are used in a certain situation. It includes device type aspects (e.g., personal computer, notebook, tablet, smartphone), hardware aspects (e.g., processor, disk space, and display size), software aspects (e.g., operating system, installed applications), and others (e.g., display properties, bandwidth).
- *Time* ( $CF_5$ ) includes time-based aspects like current time, virtual time, time zone, business days, and calendar week.
- *Environment* ( $CF_6$ ) includes environment aspects and reflects what environmental aspects influence a situation. We distinguish between physical aspects (e.g., noise level, lightening), organizational aspects (e.g., cooperate culture, enterprise policies, cooperate identity guidelines), legal requirements (e.g., privacy policy, regulations), and others.

Based on these context factors, it becomes easier to model a context. Different context modeling approaches can be used for this purpose: *key-value models*, *markup scheme models*, *graphical models*, *object-oriented models*, *logic-based models*, and *ontology-based models* [18]. In our framework, we use ontology-based models (cf. Table 2), since there exists powerful tool support for ontologies. Furthermore, partial validation and distribution of context information becomes possible and ontologies allow for an easy linking to other ontology-based models (e.g., ontology-based process information models and business process models). Finally, ontologies have strengths relating to normalization and formality. Several authors (e.g., [18]) share our assessment that ontology-based models provide a promising approach to deal with the challenge of context modeling.

**Table 2.** Comparison between context modeling approaches

Criteria	Key-Value	Markup	Graphical	Object	Logic	Ontology
Ease of use	++	+	o	o	-	o
Formalization	-	o	o	o	++	++
Expandability	--	+	o	+	-	++
Expressiveness	-	o	+	+	++	++

++: very good, +: good, o: neutral, -: bad, --: very bad

Altogether, the context model is responsible for storing context information. Based on this context information, a PAIS is able to better identify relevant process information for process participants.

## 5 Related Work

Bucher and Dinter [3] conducted a study to assess benefits, design factors, and realization approaches for POIL. Management challenges related to information logistics (IL) are discussed by Winter [19]. Heuwinkel and Deiters [20] demonstrate the possibilities and advantages of IL in the healthcare sector.

Context and context-awareness in general are discussed by Dey [7], Schilit et. al [10], and Pascoe et. al [11]. Context-awareness in IL is discussed by Haseloff [8], Meissen et. al [13], and Lundqvist et. al [21].

Further approaches have been proposed to deal with challenges of context-awareness and context modeling. Especially in the field of mobile computing a numerous of context frameworks have been proposed (e.g., Context Toolkit [22], Hydrogen [23]). More frameworks exist in the field of information retrieval (e.g., SAiMotion [24]). A broader view on context models supporting business process agility is given by Thönssen and Wolff [25].

Schilit et. al [10] investigate possible context factors. They distinguish between location, identity, and device. Dey et. al [7] state that location, identity, activity, and time are more important than other context factors. Kaltz et. al [26] propose

user & role, process & task, location, device, and time as possible categorization of web application scenarios.

## 6 Summary and Outlook

This paper proposes a context framework for enabling context-awareness in process-oriented information logistics. We motivate the need for context-awareness and show why the handling of context information is success-critical with respect to the context-aware delivery of process information. Most important, we introduce the our context framework, which deals with gathering, representing, storing, analyzing, and providing of context information along executed business processes. More specifically, we describe the framework's architecture and introduce important context factors and context aspects (to be used in context modeling).

Future research will include a more detailed investigation of the presented context aspects (also in cross-organizational scenarios), the development of a proof-of-concept application implementing our context framework, and further research on the handling of context modeling.

## References

1. Edmunds, A., Morris, A.: The Problem of Information Overload in Business Organisations: A Review of the Literature. *Int'l J. of Information Management* 20(1), 17–28 (2000)
2. Michelberger, B., Mutschler, B., Reichert, M.: Towards Process-oriented Information Logistics: Why Quality Dimensions of Process Information Matter. In: *Proc. 4th Int'l Workshop on Enterprise Modelling and Information Systems Architectures (EMISA 2011)*, Hamburg. LNI, vol. 190, pp. 107–120 (2011)
3. Bucher, T., Dinter, B.: Process Orientation of Information Logistics - An Empirical Analysis to Assess Benefits, Design Factors, and Realization Approaches. In: *Proc. 41st Annual Hawaii Int'l Conf. on System Sciences*, pp. 392–402 (2008)
4. Lechtenböcker, J.: Data warehouse schema design. Infix Akademische Verlagsgesellschaft Aka GmbH, PhD Thesis, University of Münster (2001)
5. Reichert, M., Kolb, J., Bobrik, R., Bauer, T.: Enabling Personalized Visualization of Large Business Processes through Parameterizable Views. In: *Proc. 27th ACM Symposium On Applied Computing (SAC 2012)*, 9th Enterprise Engineering Track, Trento (accepted for publication, 2012)
6. Michelberger, B., Mutschler, B., Reichert, M.: On Handling Process Information: Results from Case Studies and a Survey. In: Daniel, F., Barkaoui, K., Dustdar, S. (eds.) *BPM Workshops 2011, Part I. LNBIP*, vol. 99, pp. 333–344. Springer, Heidelberg (2012)
7. Dey, A.K.: Providing Architectural Support for Building Context-Aware Applications. PhD Thesis, Georgia Institute of Technology (2000)
8. Haseloff, S.: Context Awareness in Information Logistics. PhD Thesis, Technical University of Berlin (2005)
9. Indulska, J., Sutton, P.: Location Management in Pervasive Systems. In: *Proc. Australasian Information Security Workshop Conf. on ACSW Frontiers 2003*, Adelaide, vol. 21, pp. 143–151 (2003)

10. Schilit, B.N., Adams, N., Want, R.: Context-Aware Computing Applications. In: Proc. 1st Int'l Workshop on Mobile Computing Systems and Applications, Santa Cruz, pp. 85–90 (1994)
11. Pascoe, J., Ryan, N.S., Morse, D.R.: Human-Computer-Giraffe Interaction: HCI in the Field. In: Proc. 1st Workshop on Human Computer Interaction with Mobile Devices, GIST Technical Report G98-1 (1998)
12. McCarthy, J., Hayes, P.J.: Some Philosophical Problems from the Standpoint of Artificial Intelligence. In: Machine Intelligence, pp. 463–502. Edinburgh University Press (1969)
13. Meissen, U., Pfennigschmidt, S., Voisard, A., Wahnfried, T.: Context- and Situation-Awareness in Information Logistics. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 335–344. Springer, Heidelberg (2004)
14. Chen, H.L.: An Intelligent Broker Architecture for Pervasive Context-Aware Systems. PhD Thesis, University of Maryland (2004)
15. Baldauf, M., Dustdar, S., Rosenberg, F.: A Survey on Context-aware systems. *Int'l J. of Ad Hoc and Ubiquitous Computing* 2(4), 263–277 (2007)
16. Hipp, M., Mutschler, B., Reichert, M.: On the Context-aware, Personalized Delivery of Process Information: Viewpoints, Problems, and Requirements. In: Proc. 6th Int'l Conf. on Availability, Reliability and Security (ARES 2011), Vienna, pp. 390–397 (2011)
17. Klemke, R.: Modelling Context in Information Brokering Processes. PhD Thesis, RWTH Aachen University (2002)
18. Strang, T., Linnhoff-Popien, C.: A Context Modeling Survey. In: Workshop on Advanced Context Modelling, Reasoning and Management, UbiComp 2004 - The 6th Int'l Conf. on Ubiquitous Computing, Nottingham (2004)
19. Winter, R.: Enterprise-wide Information Logistics: Conceptual Foundations, Technology Enablers, and Management Challenges. In: Proc. 30th Int'l Conf. on Information Technology Interfaces (ITI 2008), Dubrovnik, pp. 41–50 (2008)
20. Heuwinkel, K., Deiters, W.: Information logistics, e-healthcare and trust. In: Proc. Int'l Conf. e-Society (IADIS 2003), Lisbon, vol. 2, pp. 791–794 (2003)
21. Lundqvist, M., Sandkuhl, K., Levashova, T., Smirnov, A.: Context-Driven Information Demand Analysis in Information Logistics and Decision Support Practices. In: Proc. 1st Int'l Workshop on Contexts and Ontologies: Theory, Practice and Applications (2005)
22. Salber, D., Dey, A.K., Abowd, G.D.: The Context Toolkit: Aiding the Development of Context-Enabled Applications. In: Proc. SIGCHI Conf. on Human Factors in Computing Systems: the CHI is the Limit (CHI 1999), Pittsburgh, pp. 434–441 (1999)
23. Hofer, T., Schwinger, W., Pichler, M., Leonhartsberger, G., Altmann, J., Retschitzegger, W.: Context-Awareness on Mobile Devices - the Hydrogen Approach. In: Proc. 36th Annual Hawaii Int'l Conf. on System Sciences (HICSS 2003), vol. 9, pp. 292–301 (2003)
24. Gross, T., Klemke, R.: Context Modelling for Information Retrieval - Requirements and Approaches. *IADIS Int'l J. on WWW/Internet* 1(1), 29–42 (2003)
25. Thönssen, B., Wolff, D.: A broader view on Context Models to support Business Process Agility. In: Semantic Technologies for Business and Information Systems Engineering: Concepts and Applications, pp. 337–358. IGI Global (2010)
26. Kaltz, J.W., Ziegler, J., Lohmann, S.: Context-aware Web Engineering: Modeling and Applications. *Revue d'Intelligence Artificielle* 19(3), 439–458 (2005)

# Making Recommendations for Decision Processes Based on Aggregated Decision Data Models

Razvan Petrusel and Paula Ligia Stanciu

Faculty of Economical Sciences and Business Administration, Babes-Bolyai University,  
Teodor Mihali str. 58-60, 400591 Cluj-Napoca, Romania  
{razvan.petrusel,paula.stanciu}@econ.ubbcluj.ro

**Abstract.** The decision making process is a sequence of (mostly mental) actions. But individual decision making is a fuzzy process that lacks a clear workflow structure. This issue may decrease the quality of data-centric business decisions where information must be processed in the right order and used at the right time. We argue that, when faced with such a decision, step-by-step recommendation provides help in steering the process and valuable guidance in improving it. Our Data Decision Model (DDM) is an acyclic graph that suits the fuzzy nature of decision processes. In our approach, the recommendation is based on an aggregated DDM extracted from a large number of individuals. This paper introduces two algorithms that, given a certain state of the process, provide suggestions for the next action the decision maker should perform.

**Keywords:** Decision Making Recommendation, Decision Data Model, Decision Process Mining.

## 1 Introduction

Decision making is a daily activity for every individual. Most of the decisions we make are based on experience and heuristics rather than on a scientific methodology. But what happens when a decision maker is in a decision situation in which he doesn't know where to start from or what to do next? Wouldn't it be nice if, when stuck, there was some recommendation that would indicate what others have done in a similar situation?

Our approach aims to provide the framework and tools for researching the business decision making behavior of many individuals. The decision making process is looked at as a workflow of actions directed towards choosing a decision alternative. In our previous papers we argued that it is feasible to use management simulation software that logs what users are doing while making a particular decision. We also showed how a model that aggregates the behavior of all those individuals can be mined from those logs. This paper shows how the model is used in order to provide recommendations (what to do next), given a certain state in the decision process.

We are not aiming at automating the decision process or at providing a way for choosing the "best" decision alternative out of a set of choices. Our approach is intended to provide support by steering the flow of decision maker's activities aimed at

generating and evaluating the choices. We do not aim to recommend which of the choices is the best one. Our recommended next action can either be performed or ignored by the decision maker. On the next step of the process, the recommendation is updated according to the action performed by the decision maker. This kind of support could be valuable since a decision maker may either overlook some essential aspects of a decision, or may lack the knowledge required to make a particular decision.

Our research is placed in the context of business decisions. The recommendations are based on a pre-requisite aggregate Decision Data Model (DDM). Our approach is looking at data because most business decisions are data-centric. The actions performed during a business decision process (e.g. find out the trading price of some stock; look at the trend for a stock over the last week) usually overlap with data manipulations (e.g. retrieve a figure from some source; compute the daily changes of stock prices).

The next section of the paper introduces the reader to an overview of the approach and to its formal fundamentals. The third section introduces two algorithms that can be used for providing recommendations based on a mined decision data model (DDM). In the fourth section we validate and compare the proposed algorithms using an aggregated model extracted from 50 decision makers. The last two sections deal with the related work and the conclusions.

## 2 The Approach

This first sub-section introduces an overview of our previous work in order to provide the context of the work presented in this paper. The second sub-section introduces the formal fundamentals that will be used throughout the paper.

### 2.1 The Framework

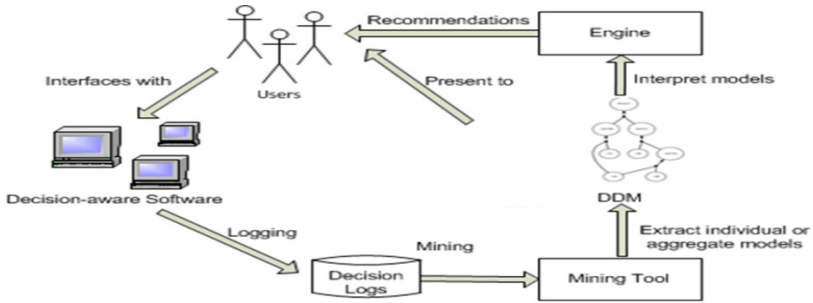
This sub-section aims to introduce the reader to the framework of decision process mining. This paper focuses on a small part of this framework (i.e. providing recommendations based on a previously mined decision model).

Everything starts (Fig. 1) with a large number of decision makers that interact with software in order to make some decision. It can be any software (even a simple spreadsheet) which simply provides the users with a set of values that are relevant to the decision to be made. The goal of the process should be to choose one of the given decision alternatives, based on the actual values shown in each scenario. Since we are interested in business decision making, we can use as examples of such software the on-line stock trading platforms, management simulation games, on-line shops, etc. One of the features of such software is that the decision maker should not be influenced in any way during the process. The software should be able to log the activity of the user by various means (e.g. following the mouse moves and/or clicks, eye-tracking, etc.). We call this kind of software 'decision-aware'. More details on the decision-aware software and on about user activity logging are available in [1].

The logs are processed by a mining application that outputs individual or aggregated Decision Data Models. More details, on how the data is mined and the individual DDM is created, are available also in [1]. The aggregation can be done for all the traces in the log or just for a selection. Once an aggregated model is created,



the process of one individual cannot be distinguished. Also, in an aggregated model, one cannot look at the most frequent path in the sense of a workflow model. But it is annotated with the frequency of the operations, so the most frequent activities of the subjects are easily visible.



**Fig. 1.** The framework of decision making process mining

Once the DDM (individual or aggregated) is created, it can be introduced as it is to the users. It can be used to either to gain a better understanding of the actions of a particular decision maker or to look at the aggregated behavior of many individuals. It is also possible to measure the distance between two DDMs (individual and/or aggregate) using different metrics. Calculating a similarity score of two users is one of the interesting outputs of our research. Another interesting output is the possibility to cluster the decision makers (based on the distance matrix of all the models).

The upper right area of Fig. 1 is where the goal of this paper lays. We show how recommendations for “best” next action to be performed, by the decision maker, can be produced. Algorithms are based on a previously extracted aggregated DDM.

The essential assumptions related to providing recommendations are:

- the decision scenario provided to the user contains all the data needed for the decision at hand. There is no other essential information relevant for the decision;
- the data provided in the scenario allows the user to explore and evaluate all the possible decision alternatives;
- a particular user might overlook certain aspects of a decision. But, all the aspects of a particular decision should be discoverable, if a large number of users are observed;
- the more frequently a certain derived data item is observed in decision making processes, the more essential it is for performing the choice of a decision alternative.

For example, our approach can be used to gain insights into buying stocks from the stock market. We could add the logging feature to a site providing stock information (e.g. <http://www.reuters.com/finance/stocks>). Then, we could ask a lot of users to start deciding whether to buy or sell stock and log their actions while making the decision (data that is looked at, data that is compared or added, etc.). Based on the aggregated DDM we could provide recommendations about which data is important. We can suggest the order of finding out and using it based on frequent elements of the DDM.

In order to improve the understanding of the aggregate DDM and the algorithms based on it, we will use a running example. Let’s suppose that we observed 100

decision makers while making the same decision, based on the same data. The sequences of operations and their frequency (F) are: F 10 (opA, opB, opC, opD, opB, op1, op2, op3), F 10 (opA, opB, opC, opD, opB, op1, op2, op3, op4), F15 (opA, opB, opC, opD, op1, op2, op5), F15 (opA, opB, opC, opD, opE, op1, op2, op5, op6), F20 (opA, opB, opC, opD, opF, op1, op2, op5, op7), F30 (opC, opD, opA, opB, opF, op2, op8). Starting from those traces, we can mine the aggregated DDM (in Fig. 2). Each operation is annotated with the frequency e.g. opA shows up 100 times.

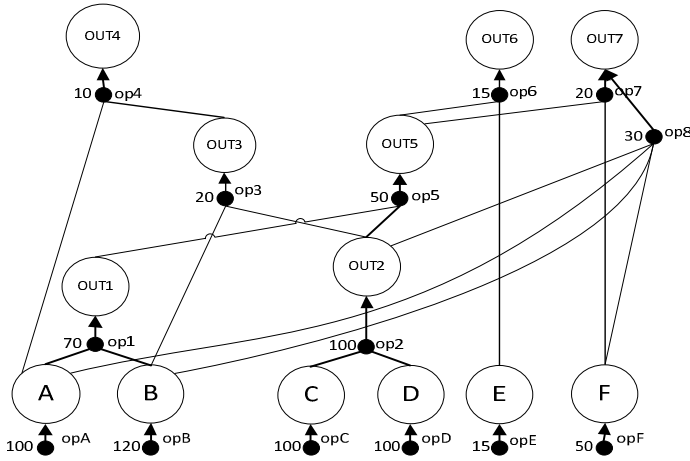


Fig. 2. A Decision Data Model used as the running example

The operations are labeled according to the type of output (i.e. if the input is the empty set then it produces a basic data items and is labeled with its name; or else it produces a derived data items and is labeled artificially). The semantics of the model implies that an operation can be executed only if its input data items are known (e.g. op1 can be executed only if the values of A and B are both known). If we use the example with the stock market, the model states that the trend of some stock (out1) can be calculated only if the previous price (A) and the current price (B) are known.

## 2.2 The Formal Approach

This sub-section introduces the essential notions used further in the paper. Basically, we are looking at a search problem in an acyclic, not rooted graph (DDM). The problem of providing recommendations is to determine the “best” path through the DDM. Our problem can be mapped to a general search problem with five components:  $S$ ,  $S_0$ ,  $S_g$ , *successors* and *cost* (where  $S$  is a finite set of states;  $S_0 \subseteq S$  is the non-empty set of start states;  $S_g \subseteq S$  is the non-empty set of goal states; *successors* is a function  $S \rightarrow P(S)$  which takes a state as input and returns another state as output (probabilities may be used in connection with it); and *cost* is a value associated to moving from state  $s \in S$  to  $s' \in S$ ). The total cost is the sum of the *costs* incurred by a sequence of movements from state  $s \in S_0$  to a state  $s' \in S_g$ . A recommended strategy is a sequence of actions such as the total cost is minimized (or

maximized under some circumstances). A particular feature of our problem is that, because we use simulation software, there are no costs for moving from a state to the next (as used in classical search problems). Instead, the notion of cost is derived from the notion of frequency.

*Definition 1:* A Decision Data Model (DDM) is a tuple  $(D, O)$  with:

- $D$ : the set of data elements  $d$ ,  $D = BD \cup DD$  where  $BD$  is the set of basic data elements and  $DD$  is the set of derived data elements;
- $O$ : the set of operations on the data elements. Each operation,  $o$  is a tuple  $(d, v, DS, t)$ , where:
  - $d \in DD$ ,  $d$  is the name of the output element of the operation;
  - $v$  is the value outputted by the operation. Can be numeric or Boolean;
  - $DS$ , a set of  $d \in D$ , is the set of input data elements of the operation.
  - $t \in T$ , where  $T$  is the set of timestamps at which an operation from  $O$  occurs (i.e. the time when the element  $d$  is created using  $o$ ).
- $D$  and  $O$  form a hyper-graph  $H = (D, O)$ , connected and acyclic.

An example of a DDM is depicted in Fig. 2.; where  $BD = \{A, B, C, D, E, F\}$ ,  $DD = \{Out1, Out2, Out3, Out4, Out5, Out6, Out7\}$  and  $O = \{opA, opB, opC, opD, opE, opF, op1, op2, op3, op4, op5, op6, op7, op8\}$ . For example, the operation  $opA = (A, 100, \emptyset, time\_10)$  and  $op3 = (Out3, 1000, \{B, Out2\}, time\_100)$ .

*Definition 2:*  $o_i = (d_i, v_i, DS_i, t_i)$  and  $o_j = (d_j, v_j, DS_j, t_j)$  are *alternate (or mutually exclusive) operations* if  $d_i \neq d_j$  and  $v_i \neq v_j$ . One can note that, in Fig. 2,  $op7$  and  $op8$  output the same derived data item ( $Out7$ ). However, they are not the same operation because their inputs are different. Mutually exclusive operations are important because if, for example,  $op7$  is performed there is no need to perform  $op8$  and the other way around.

*Definition 3:* An *aggregated DDM* is a DDM annotated with the *frequency* of the operations. *Frequency* indicates how many times an operation shows up in the log. In Fig. 2 the number near an operation indicates its frequency.

*Property 1:* An aggregated DDM allows the user to ‘zoom-in’ or ‘zoom-out’. Zooming-in shows less frequent behavior while zooming-out shows only operations with a frequency above a certain threshold. This is extremely useful when the aggregated DDM shows a lot of outlying behavior. For example, the outliers can be removed by setting the threshold value to 2 (i.e. the same operation should be performed by at least two different decision makers).

*Definition 4:* An operation  $o = (d, v, DS, t)$  is *enabled* when the input data elements ( $DS$ ) are available (known to the decision maker). If an operation is enabled it may be executed so that the output element  $d$  is produced and the output value  $v$  is known. When any operation is executed, the process moves from one state to another.

*Property 2:* All the basic data elements have as input the empty set. Therefore, at the start of the decision process, the only enabled operations are the ones producing the basic data elements. Executing such operations may be seen as the stage in any decision process in which the decision maker finds out the specific data.

*Definition 5:* A *final derived data element* is a data item that is not used as input in any operation. We assume that it is a criterion for performing the choice and therefore, given a rational decision maker, it is a sub-goal of the decision process.

*Definition 6:* A *state of a DDM* is a particular distribution of operations over the sets of Enabled, Not-Enabled and Executed operations. For example, the initial state is the one in which all the operations that produce the basic data elements are enabled, all the operations producing derived data elements are not-enabled, and there is no executed operation (e.g. for the DDM in Fig. 2, Enabled = {opA, opB, opC, opD, opE, opF}, Not-Enabled = {op1, op2, op3, op4, op5, op6, op7, op8} and Executed =  $\emptyset$ ). As the process progresses, the different states are represented by different placements of the operations in the three sets. The end state is reached when all operations are placed in Executed set, while Enabled and Not-Enabled are empty. There is no need to reach the end state in order to make the decision.

### 3 The Recommendation Algorithms

This section introduces the algorithms that produce recommendations for a decision process, based on a DDM. We aim to provide the answer to one question: “Which is the action that the decision maker should do next, given his previous actions?”. Therefore, the problem we are faced with can be stated such as: “How to select the best operation to be performed next (from a set of enabled operations), in any given state of the process?”. As explained before, the state of the process is a certain distribution of the DDM operations over the Enabled, Not-Enabled and Executed sets. The decision maker can choose to execute any operation from the Enabled set, or may perform any new operation that is not in the model (as long as the data needed as input for that operation is available). If an Enabled operation in the DDM is executed, it is moved to the set of Executed operations. After the decision maker performs an operation (the suggested one or any other), the process moves to a new state and the system provides another recommendation.

We assume that:

- each operation can be executed only once (since it is useless to calculate something once you know its value);
- each operation can only be executed successfully;
- the more frequent an operation is performed, the more important it is. Therefore the primary goal of any algorithm should be to identify the most frequent path in the model.

There are several possible approaches over DDM based recommendations, according to the underlying assumptions. For example, we can either assume the decision process has memory or is memory less (e.g. moving from one state to the next one depends on all previous states or only on the current state). Or, we can assume the process can take infinite time or is restricted to a given time horizon. Therefore, we developed several algorithms. Given the space limits we can introduce only two in this paper, each looking at the problem from two perspectives. The naïve one suggests the next operation by considering the absolute frequency. It has no clear target, and only aims to guide the user through the most frequent operations. The second algorithm assigns priorities to operations producing a final derived data element (which are actually the decision

criteria). Then, it guides the user along a path so that the operation producing a certain final data element is reached at a minimal cost.

*Algorithm 1:* We first introduce a naive algorithm which uses a Greedy approach, recommending the most frequent operations that is enabled.

1. Let  $DDM_{agg} = (D_{agg}, O_{agg})$ ;
2. Let  $op$  be the list with the operations in  $O_{agg}$ ;
3. Let  $no\_of\_occurrences$  be the list with the number of occurrences for each  $op$ ;
4. Select  $op$  with  $\max(no\_of\_occurrences)$  and place it in  $Max\_Occ$  set;
5. Compute *Enabled* and *Executed* sets;
6. For each in *Executed* set, search for mutually exclusive operation. If found, move them from *Enabled* set to *Executed* set;
7. Compute  $Recommendation = Max\_Occ \cap Enabled$

**Table 1.** Example of recommendation for the running example using Algorithm 1

State	Enabled	Not enabled	Max Occ (frequency)	Reco mm	Executed
1	opA, opB, opC, opD, opE, opF	op1, op2, op3, op4, op5, op6, op7, op8	opB (120)	opB	∅
8	opE, op3, op5, op8	op4, op6, op7	op5 (50)	op5	opB, opC, opD, op2, opA, op1, opF
9	opE, op4, op5, op8	op6, op7	op5 (50)	op5	opB, opC, opD, op2, opA, op1, opF, op3
10	opE, op4	op6	op5 (50)	op5	opB, opC, opD, op2, opA, op1, opF, op3, op8, op7
12	op6, op4	∅	op6 (15)	op6	opB, opC, opD, op2, opA, op1, opF, op3, op8, op7, op5, opE,

In the initial state none of the data items is known and the Enabled set contains the operations that produce the basic data items, while all the operations producing derived data items are Not-Enabled. In this initial state, the most frequent enabled operation (opB) is recommended. Later in the decision process (State 8), op5 is recommended but user decides to perform op3 (see last item in Executed set in State 9). The system again recommends op5 since it is still the most frequent one enabled. In State 9, again the user ignores the recommendation and performs op8. Performing op7 becomes pointless, so it is moved to Executed set (see last two items in Executed set in State 10), even if it wasn't enabled. The reason is that the user already found out the value of the derived data item out7, therefore there is no need to calculate it again. The user decides to make the decision without performing all the operations (see State 12 where there are still two Enabled operations, while the performed ones are logged in Executed set). The DDM will be updated according to the new trace, so the frequency for some operations increases by 1, while for others remains the same.

*Algorithm 2:* This approach is inspired from the A\* path finding algorithm [2].

1. Create array *Final* with the operations that output final data elements ( $fo$ ) and their frequency ( $f_{fo}$ );
2. Use depth-first search to calculate the direct paths to each element in *Final* and place them in *Paths*;

3. Evaluate each in *Paths* using formula  $F_i = (G + H)$  where  $F$  is the score of each path,  $G$  is the total individual cost of the operations executed in the prior states of the process and  $H$  is the total cost of the remaining operations along the selected path. The cost of an operation is calculated as the sum of the frequencies of all operations divided to the frequency of that operation;
4. *Current path* = the element from *Paths* where  $F_i$  is minimal;
5. *Recommendation* = max frequency ( $Enabled \cap Current Path$ );
6. If *Recommendation* =  $\emptyset$   
     End  
 Else Compute New State and go to step 3.

In Fig. 2, the ranked list of pairs comprising operations leading to final derived data and their frequencies is  $\langle (Op8, 30), (Op7, 20), (Op6, 15), (Op4, 10) \rangle$ . There is only one path that leads to enabling op8 and can be determined using depth-search as {op2, opC, opD, opA, opB, opF}. There is also only one path to op6 which is {op5, op1, opA, opB, op2, opC, opD, opE}. In the initial state (there are no operations in Executed) the first objective would be to perform op8 because  $F_{op8} = (0 + 81.34)$  while  $F_{op6} = (0 + 152.75)$ . In Table 3 we show State 6 in which  $F_{op8} = (50.09 + 42.66) = 92.75$ . The total score increased because the user decided to perform op1 which is outside the path to op8. However,  $F_{op6}$  is still 152.75 because the executed op1 was in its path. In State 8 the objective has changed because  $F_{op6} = (99.41 + 53.33)$  and the user only needs to perform op6. Meanwhile, to perform op8 the user needs to perform first opF and then op8 at a higher total cost.

**Table 2.** Example of recommendation for the running example using Algorithm 2

State	Enabled	Not enabled	Objective (score)	Reco mm	Executed
6	opE, opF, op3, op5	op4, op6, op7, op8	op8 (92.75)	opF	opA, opB, opC, opD, op1, op2
8	opF, op3, op6	op4, op7, op8	op6 (152.75)	op6	opA, opB, opC, opD, op1, op2, opE, op5

One can notice that there is a potential problem with the Greedy approach (Algorithm 1). It can get stuck in providing the same recommendation over and over if there is a high frequency operation that is repeatedly ignored by the user. The second algorithm adapts itself to the decision process by changing the objective if a path with a lower cost is available to a final derived data element (decision criterion) when the user repeatedly ignores the recommendation.

## 4 Case Study

This section aims to provide a comparative evaluation of the algorithms introduced in the previous section. The evaluation uses the decision-aware implementation we created (available at [www.edirector.ro/v3\\_1](http://www.edirector.ro/v3_1) and accessible with username and password “test”). The decision problem to be solved by the subjects is to decide if they buy or rent a house. Some of the data elements available are the price of the house, the savings, the monthly income, etc.

The entire log we use in this section can be downloaded from [http://www.edirector.ro/v3\\_1/export/pm.xml](http://www.edirector.ro/v3_1/export/pm.xml). From a selection of 50 user traces we created an aggregated DDM that was used as input to the recommendation algorithms. The aggregated DDM was ‘zoomed out’ (i.e. we selected only elements with a frequency at least 3) to get rid of the exceptional behavior.

We set up an experiment involving a focus group of 9 decision makers (users). Since the number of subjects is small for establishing a proper scientific claim, we will replicate this experiment. The objective of the experiment is two fold:

- to find out if there is any difference between the unsupervised decision making process and the one in which some recommendations are provided;
- to find out which of the algorithms produces recommendations used more frequently by the decision makers.

The experiment is divided into two parts. First, each subject is required to make a decision given the scenario data, without any recommendation. Then, the subject is provided at each step with recommendations, as generated by each algorithm. Each usage of the software produces a trace which is actually a sequence of actions. We also logged the sequence of actions recommended by each algorithm. At the end of the experiment we conduct an interview with the subjects for a qualitative assessment of the experience with the software and the recommendations.

The measured variable relevant for the first sub-objective is the distance between traces. The metric we used for trace comparison is Jaccard index. The similarity score of two traces, A and B,  $score(A,B)$  is calculated as the number of identical operations ( $A \cap B$ ) divided to the union of operations in the two traces ( $A \cup B$ ). This score abstracts from the sequence of operations.

The measured variable for the second sub-objective is the recommendation number of ‘hits’ for each algorithm (i.e. how many times the user followed the recommendation). We cannot use precision and recall as used in information retrieval since there is no notion of ‘good’ or ‘bad’ recommendation. The ‘hit’ can be substituted for the notion of ‘good’ recommendation but we argue that the user is biased since the recommendation is disclosed before he makes his choice. Even more, the direct observations during the experiment revealed that, if the first few actions were hits, the user begins to trust the recommendation and follows it more often.

There are several risks that threaten the validity of the experiment. One concern is the quality of the aggregated DDM. It was mined mostly from the traces of master students at our business faculty. Therefore, we see it as halfway between expert and beginner. A second concern is the subjects of the experiment. We used 3 expert (i.e. advanced knowledge of the decision and lots of experience with the software), 3 intermediate (regular knowledge of the decision and some previous experience with the software) and 3 beginner subjects (regular knowledge of the decision and first time users of the software). A third concern relates to the degree of the software’s interface influence over the subjects. It can be easily observed that most of the traces start by evaluating the purchasing data simply because this is the first tab. Even more, the layout of the textboxes is important (e.g. the majority of traces start with op1 and then follows op2 because these are the first textboxes of the first tab). To mitigate this risk, for the next experiments we will change the order of the tabs and also reorder the textboxes in each tab. And finally, since the number of subjects is not statistically

relevant, we used the simple average. As the experiment will be repeated and more data gathered, we will employ more advanced metrics.

Because of the limited space, in Table 4 we show just one trace for each category of subjects. From the log introduced at the beginning of the section, the unsupervised (supervised) traces for experts are 181 (195), 183 (194), 217 (218), for intermediates are 197 (198), 201 (207), 213 (215), and for beginners are 185 (187), 190 (192), 211 (214).

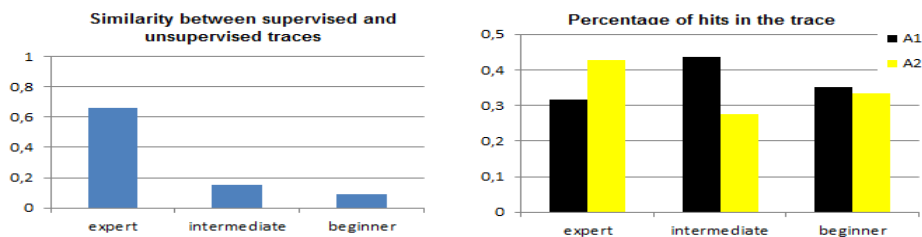
**Table 3.** Traces for an expert (user 1), intermediate (user 2) and beginner (user 3)

User	Unsupervised trace	Supervised trace	Recommendation A1	Recommendation A2
1	op1, op2, op29, op3, op4, op5, op50, op14, op6, op10, op16, op51	op1, op2, op14, op29, op5, op3, op4, op52, op6, op10, op16, op15, op53, op54, op55	op1, op2, op6, op6, op6, op6, op6, op6, op6, op5, op7, op27, op27, op6, op11, op11, op11, op11, op11, op11	op14, op14, op14, op29, op5, op7, op27, op27, op27, op10, op16, op15, op7, op7, op7
2	op1, op2, op26, op28, op63, op7, op24, op64	op1, op2, op29, op5, op3, op4, op6	op1, op2, op6, op6, op6, op6, op6	op14, op14, op29, op5, op7, op7, op27
3	op2, op65, op66, op67, op68, op14	op1, op2, op14, op29, op5, op7, op24, op13, op3, op4, op27, op26, op6, op11, op5, op20	op1, op2, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6, op6	op14, op14, op14, op29, op5, op7, op24, op13, op3, op4, op27, op26, op28, op28, op10, op28

Examining the data in Table 4 one can notice the main limitation of Algorithm 1 concerning the re-occurrence of the same recommendation. On the other hand, Algorithm 2 adapts more to the actual process being performed.

The Jaccard distance calculated between the unsupervised and supervised traces for user 1 is 0.69. This reveals that the user changed his process because of the recommendations. The number of hits for Algorithm 1 is 6 while the number of hits for Algorithm 2 is 12, therefore we can argue that the user found the recommendations of Algorithm 2 more useful.

From the first chart in Fig. 3 we can answer the first research question and safely argue that the recommendation had a large impact on the decision process for all types of users. The average change is close to 40% for the experts and it is dramatic for the intermediates and the beginners (more than 80% of the actions were different).



**Fig. 3.** Experiment results

Considering the second chart in Fig. 3 we see that the experts preferred the recommendations of Algorithm 2 (more than 40% of the operations performed were the ones recommended by A2). Interestingly, the situation is inverted for the intermediates. The beginners seem to prefer both recommendation algorithms. But,



the main point of this chart is that none of the processes follows closely one recommendation. We conclude that the subjects did use the recommendations, but didn't rely only on it. This chart doesn't provide a definitive answer to the second research question. It seems that the adaptive Algorithm 2 follows more closely the more structured process of experts. Meanwhile, the 'dumber' Algorithm 1 seems to work better with individuals that have some insight into the decision process but not a clearly structured process in mind. We will further investigate this issue.

## 5 Related Work

The work introduced in this paper may seem somewhat similar to providing recommendations in web-based systems. There, too, is a large log with user actions (e.g. what items were purchased) based on which some recommendations are provided. But the fundamental approach focuses on the associations between properties of items. We do not focus on the choice itself but on the reasoning process that leads to that choice. Therefore, the work done in content-based or collaborative filtering poorly fits our needs.

From artificial intelligence field there are many algorithms that search graphs [2]. We found that A\* algorithm fits our problem and adapted it for use in Algorithm 2.

The decision processes can be represented, and recommendations can be produced, using some executable workflow formalism (e.g. Petri Nets, BPEL) [3]. But, in those approaches, adding behavior to a model is not a trivial problem. Our entire approach has at its core the idea of extracting a model from a large number of individuals. Therefore, as new individuals use the software, new behavior is logged all the time. A DDM is easily and fast updated [1], therefore the recommendation algorithms performance is not reduced. Even more, the workflow approaches focus mostly on the control flow perspective [3]. Instead, we focus on the data flow perspective aimed at producing a choice. Considering this focus, we found our inspiration in the Product Based Workflow Modeling approach [4]. The approach has at its core a model called Product Data Model (PDM). The DDM is derived from the PDM, having some specific features and properties.

Given a PDM, providing recommendations of the next action to be taken can be done using Markov Decision Process (MDP) approach [5]. It is suggested that the MDP-based recommendation is difficult to apply for real situations because of the State Explosion problem. It is argued that it can be partly avoided by applying heuristics (i.e. by replacing the MDP global choice with local choices). A DDM-specific property is that an operation can only be executed successfully, and the order of operations for basic data is irrelevant. Therefore, the state space stays small.

Applying decision trees to our approach is unfit since we look at decision making as at a workflow and we don't aim to classify the actions of the user. Even more, applying decision trees to such an approach would lead to a state explosion problem with respect to the size of the tree.

Our evaluation of the algorithms was inspired by [6]. The offline analysis is suggested to be used mainly in evaluating predictive accuracy. Our goal is not to predict what a certain decision maker would do next. On the other hand, live experiments fit reactive recommendation systems. Since all our recommender algorithms are reactive, we used a focus-group based experiment.

## 6 Conclusions

The research presented in this paper is placed in a data-centric business decision making context. The basic aim of this paper is to introduce several algorithms that allow an individual to perform a guided walk through a particular type of acyclic graph. The graph is extracted from a log containing the activities of large numbers of persons performed while making a specific decision.

We created two algorithms: one that uses a Greedy approach and one that looks at the previous actions in the process and determines the best path towards a sub-objective of the process (a decision criterion).

Our evaluation shows that providing a recommendation changes the decision process for all classes of users (experts, intermediates or beginners). However, our experiment didn't allow us to state, at this point, which of the two algorithms is better.

We are aware that we didn't cover all possible types of algorithms. For example, we are developing a mapping of our problem to a Markov Decision Process since it may produce better results than the algorithms presented in this paper.

The future work will also aim to create clusters of similar DDM and dynamically place the user in such a cluster. The recommendation algorithms will be applied on the clustered aggregated DDM rather than on the entire aggregated DDM. Therefore, we will be able to provide better context-aware recommendations.

**Acknowledgments.** This research was supported by Human Resources Development Operational Program through the project Transnational Network of Integrated Postdoctoral Research in the Field of Science Communication, Capacity Building (Post-doctoral School) and Scholarship Program (CommScie) POSDRU/89/1.5/S/63663.

## References

1. Petrusel, R., Vanderfeesten, I., Dolean, C.C., Mican, D.: Making Decision Process Knowledge Explicit Using the Decision Data Model. In: Abramowicz, W. (ed.) BIS 2011. LNBI, vol. 87, pp. 172–184. Springer, Heidelberg (2011)
2. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Upper Saddle River (2003)
3. van der Aalst, W.M.P.: Process Mining. Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg (2011)
4. Reijers, H.A., Limam, S., van der Aalst, W.M.P.: Product-based Workflow Design. *J. of Management Information Systems* 20, 229–262 (2003)
5. Vanderfeesten, I.T.P., Reijers, H.A., van der Aalst, W.M.P.: Product-based Workflow Support. *J. Information Systems* 36, 517–535 (2011)
6. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 5–53 (2004)

# Investigation of Performed User Activities in Overall Context with IT Analytical Framework

František Babič, Jozef Wagner, and Ján Paralič

Department of Cybernetics and Artificial Intelligence,  
Faculty of Electrical Engineering and Informatics, Technical university of Košice,  
Letná 9/B, Košice, 042 01 Košice, Slovakia  
{frantisek.babic,jozef.wagner,jan.paralic}@tuke.sk

**Abstract.** Collaborative user activities represent an important source of knowledge and experiences that can be identified with suitable discovery methods. These activities are typically realized within collaborative environments which in several cases offer only a simple analytical module providing basic statistical information. However, these results are not sufficient for deeper understanding of performed activities, processes behind realized actions, hidden relations, etc. A motivation to discover and understand such knowledge from user activities led us to design a formal representation of historical events and method for their advanced analysis. This paper describes and evaluates a semi-automatic procedure of pattern definition and discovery that can assist and simplify manual evaluation of collected data by teachers, researchers or evaluators.

**Keywords:** user activities, event log patterns, analytical tool, knowledge discovery.

## 1 Introduction

Collaborative activities are driven by various elements and models of user behavior. This behavior represents an interaction between different users or between users and an environment. Its aim is to reach specified objectives in an effective and successful way. The decision process of determining which iteration or subset of realized activities is the most effective or most successful one is a very complex task and it needs to be supported by event logs. This complex evaluation can be performed based on a collected historical data representing investigated activities. We have identified several possible methods and approaches to analyze this kind of data (as described in Related work section), but our primary motivation was to automate a process of identification of patterns or knowledge creation processes in a large set of data, which is currently performed manually by end users. Typical example of this procedure can be found in collaborative learning environments. Such environment provides a virtual space for various learning courses having students which produce different outputs during their activities. When course finishes, it is necessary to evaluate used

procedures and methods, taken decisions, created outputs, etc. A teacher or evaluator has three possibilities of how to get the expected assessment: he can use his notes created during the whole course; export and manually analyze data from learning environment within e.g. an Excel; or use some analytical tool (included in a learning environment, or from a third party). The last two options represent an interesting approach on how to process a huge set of data, but it often provides only limited possibilities of preprocessing and query customization. The generation of association rules can be understood as a typical representative of this approach in which most frequent actions or objects are aggregated into rules in form of if X and Y, then Z.

Our motivation was to provide a simple and intuitive visualization of performed user activities in which users have a possibility to more easily identify activity paths which led to an important point in a process, important decision or results. The proposed solution is oriented on collaborative processes within learning environment with no predefined model. This contrasts with business processes and analyses within field of process mining, which we also describe in the Related work section. In many cases the collaborative processes result in the creation of a new knowledge. This knowledge creation process is not known beforehand and it is often unique, so it is necessary to evaluate which procedure or sequence of performed steps led to most valuable knowledge.

This approach can be also adapted to some other systems as virtual learning environment. Interesting application case represents a software testing environment that provides all necessary features for effective realization of testing procedures. Typical software testing consists of relatively well structured processes, in which we can identify many similar procedures, decisions and steps' sequences. These findings can lead to continuous improvement of relevant testing practices in terms of cost reduction, shorter execution time and more effective management of human resources. This is objective of our cooperation with software company called RWE-IT (see subsection 4.4).

Identified process aspects represent one possible view on performed user activities which can be extended with other methods as e.g. text mining analyses on relevant text documents. For this purpose we designed a text mining library called JBOWL in order to develop own implementation of relevant algorithms based on available services. More about this approach is described in subsection 3.3 below.

The methods presented in this paper represent a joint work of research team containing members from two research areas at the Technical university of Košice: Artificial Intelligence and Business Information Systems. This research team offers a good collective space to share ideas and experiences within existing or new education programs, research and project activities.

The whole paper is organized into following main sections: introduction with motivation (this section 1); the presentation of related work (in section 2) with some findings for our research, the technical description of proposed framework (section 3); four application cases in section 4 that are divided into two categories (finished and planned) and a short summary which concludes the paper (section 5).

## 2 Related Work

The core of our framework lies in its analytical capabilities; therefore we focus this section on several relevant research directions which can be identified in current state of the art research.

We start with some specific areas of data mining research, relevant to our framework. As our primary role is education and many of the processes in the knowledge society deal with some form of learning, we have analyzed the area of educational data mining, which covers the exploitation of data from educational systems or environments [2]. It deals often with temporal, noisy or incomplete datasets [13]. Other important area is process mining, which usually concentrates on business processes and provides extraction of different useful information from event logs, such as actual version of process model, comparison with its planned version and identification of deviations, as well as provision of various performance statistics, time overview of activities, or social network analyses [15].

Complex event processing (CEP, also known as Event Stream Processing) represents an interesting new approach to analysis, providing techniques for processing large amounts of event logs, often in real time. These types of analysis are heavily tied to business processes, and are used e.g. in analysis and monitoring of financial transactions, stock markets and data from RFID chips. Most active research in this field is in SASE<sup>1</sup> (*Stream-based And Shared Event processing*), ETALIS<sup>2</sup> (*Event-driven Transaction Logic Inference System*) and Cayuga<sup>3</sup>. Recently released open source library called Storm<sup>4</sup>, developed by a Twitter, Inc., supports many of CEP techniques and can be used on diverse types of event logs.

The interactions investigation in virtual environment by means of Social Network Analyses is described in [8], [10] or [14]. All these approaches are based on collection of historical logs representing performed activities and offer possibilities to identify social structures, hidden interactions and relations, etc.

The PANdit [6] "Pattern Analysis and Discovery Tool" works with user activities stored in custom log format and searches through them according to user defined patterns. These patterns are represented as rules and are created with the help of the tool's user interface. With this so called "interaction analysis" user can search for occurrences of various groups of events, or create a nontrivial filter to select interesting events. Searching is implemented in Prolog language, and results are presented on a time line as dots or poly lines.

The exploitation of suitable data mining methods to improve the testing management is described in [9] and [7]. Both examples are focused on reducing the testing time by reducing the number of test cases with identification of similar patterns or the most probably attributes of software behavior.

---

<sup>1</sup> <http://avid.cs.umass.edu/sase/index.php?page=home>

<sup>2</sup> <http://code.google.com/p/etalis/>

<sup>3</sup> <http://www.cs.cornell.edu/bigreddata/cayuga/>

<sup>4</sup> <https://github.com/nathanmarz/storm>

The aim of this section is to provide some theoretical overview about existing approaches or methods with similar motivation and focus on activity. All of these activities are based on collected data in different formats of logs. We have investigated several types of logs to identify advantages and disadvantages of each solution. Based on these findings we proposed a generic format that provides a rich data structure for analytical purposes. The traditional data mining methods with some extensions (education and process mining) in combination with requirements from teachers and researchers inspired us to define an analytical framework containing all necessary services described in following sections. The main aim is to provide a tool that enables deeper understanding of performed user activities by means of simple data presentation and management from the users' perspective.

### **3 Proposed Analytical Framework**

The proposed technical solution for analytical framework is a combination of existing services (developed mainly within the integrated FP6 project KP-Lab) and newly designed and implemented services, based on identified requirements within selected application cases. Moreover, existing services are being adapted for the new conditions in order to fulfill the specific goals of particular cases.

The core of our solution contains services for event logging, logs storage, manipulation with logs, extraction and visualization based on different user queries [12]. Extracted and visualized information represent a complex view of user behavior during virtual activities or processes, e.g. timeline-based visualization, quantitative statistics, level of collaboration, tacit relations, patterns, etc. The successful realization of all these approaches depends on quality of collected historical data. For these purposes, a generic log format was designed containing all necessary information. The initial list of 12 parameters can be simply expanded if necessary. It contains information about time, object of interest, type of action, actor performing the action and arbitrary custom data. The detailed description is presented in [12].

#### **3.1 Activity Paths and Patterns**

Traditional approaches to the analysis of performed user activities, such as process and data mining, require that logged events contain information about process instances to which the event belongs. This information is then used for process discovery and modeling. If no model of a process and no explicit information about process instances are given, traditional analyses are of little use. However, in collaborative processes (research and innovative education processes) most of the time this is the case, as historical data about user activities often contain traces of unique, ill defined processes, which are of great value for the researchers.

For such purposes, we have developed and implemented an analytical methodology and tools for analysis of so called patterns, which identify parts of processes captured in logs. Analysis is interactive, iterative and consists of following steps:

1. Understanding of problem's domain, formulation of hypothesis
2. Acquiring logs of users' actions and basic understanding of them
3. Preprocessing and creation of a filter in order to select and prepare suitable data set for analysis
4. Creation of a pattern and its parts.
5. Performing search for a pattern occurrences in given data set
6. Interpretation of results, iteration (*back to step 3 or 4*)

Pattern is a collection (usually a sequence) of fragments, each describing a generalization of some activity. Searching finds occurrences of a given pattern in a data set. Fragments are matched with events, resulting in a search tree. Results are then represented as leaf nodes in a lowest level. Process of searching with a following pattern (uppercase symbols denote variables which bind to concrete values from event logs) is shown in Fig. 1:

```
(def f1 {:actor :X :type "opening" :entity "doc1"})
(def f2 {:actor :X :type "creation" :entity :Y})
(def f3 {:actor :X :type "link" :entity :Y :link-to "doc1"})
(def pattern [f1 f2 f3])
(search data pattern)
```

Above pattern finds situations where user (any user X) created a new document (Y) after reading particular existing document (doc1), and then linked the two (previously existing document doc1 with the newly created document Y) together. In our application case the search found two results (user a2 created documents doc4 and doc6 and linked both of them to doc1), depicted as green leaf nodes. Searching process operates with two variables, X and Y, defined in the patterns above, which are bound to values as the search progresses.

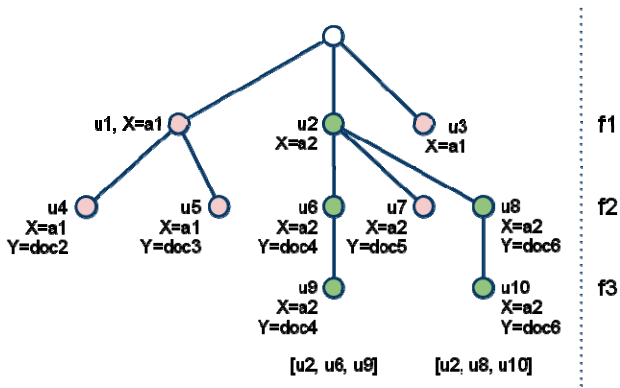


Fig. 1. Search tree for a pattern analysis

Depth-first search is used, with some optimizations which remembers environment (current fragment and variable bindings) of traversed nodes and does not expand new nodes if they happen to have same environments as the failed ones (note that node u3 does not expand).

Analysis was implemented within library called Piaget, in a functional language called Clojure, running on top of JVM (Java Virtual Machine). It provides an API for log import, preprocessing, pattern definition and searching. For the experiments, we have developed a simple user interface for both browsing and viewing of event logs and for pattern definition and search. Tools allow for creation of a pattern either from a scratch (step-by-step definition of all its parts), or by generalization of selected set of events from a log.

### 3.2 Quantitative Analyses

Quantitative analyses were designed to offer basic summarized information about performed user activities. Similar feature is available in most collaborative environments (like e.g. Moodle), but our aim was to provide unified middleware services on top of the logs which can be used by various graphical user interface (GUI) tools fulfilling different kinds of users' requirements. Such a separation of middleware analytic services and GUI services is not offered by any of current collaborative environments and presents our original contribution to the state-of-the-art providing higher flexibility for analytical as well as GUI services.

The data extraction services are implemented on the middleware level and a simple example of aggregation service is following: String *activityAggregation* (Query *query*, List<AggregationFunction> *aggregationFunctions*, Set<GroupBy> *groupBy*), where:

- *Query* parameter describes constraints used for filtering of the actions included in the aggregated view. Query object encapsulates the parameters from log format and two additional constraints for more specific description, as *filter* - set of key value pairs which will be compared with events custom properties, *excludeFilter* - true or false, whether include or exclude events which do not have properties from the filter present in them.
- *AggregationFunctions* specify the list of aggregation functions included in the view computed from the set of selected events as *NumOfActivities*; *NumOfActors*, *NumOfEntities*, *TimeSpan* - starting and ending date of investigated time period.
- *GroupBy* specify clause for the grouping of the result by actor, object or type of action.

The results depend on type of investigated environment, processes or activities, but it could be e.g. as number of participants involved and number of actions performed by each of them; number of shared objects used / changes made / versions produced; number of comments added; number of meetings, links, etc. in given time interval, within given group or with other constraints posed by the user in the analysis phase. The detailed description can be found in [11]. On the other hand we have starting a development process of our own GUI for quantitative analyses to provide simple accessible web interface with possibilities to define required queries and visualize the obtained results within friendly and understandable graphical format. This GUI will complement our analytical framework at the end of this year



### 3.3 Text Mining Services

Important source of contextual information is not presented in the log files, but rather in textual form (reports, working papers, discussion contributions, etc). Therefore one type of analytical services in our framework is based on text mining techniques. For this purpose we utilized our JBOWL<sup>5</sup>. JBOWL is an open source Java library that was designed to support different phases of the text mining process and offers a wide range of relevant classification and clustering algorithms. Its architecture integrates several external components as JSR 173 – API for XML parsing or Jakarta Lucene for indexing and searching. The motivation behind the design of this library seven years ago [3] was existence of many partial implementations of different algorithms for processing, analyses and mining in text documents within our research team on one hand side and lack of equivalent open source tools at that time. The main aim was not to provide simple graphical user interface with possibility to launch selected procedures but to offer set of services necessary to create the own text mining stream customized to concrete conditions and specified objectives. At the time of beginning (2005) several similar alternative existed as Lucene, GATE and Weka, but each of these systems covered only limited range of features providing by JBOWL [3].

The initial set of JBOWL functions:

- management and manipulation of large sets of text documents,
- support for different formats as plain text, HTML or XML in both languages: Slovak and English,
- services for indexing, complex statistical text analyses and preprocessing tasks,
- interface for knowledge structures as ontologies, controlled vocabularies or database WordNet

has been continuously extended and improved based on new requirements or expectations expressed by researchers and students of our department. The last JBOWL update offers possibility to run the text mining tasks in a distributed environment within task-based execution engine. This engine provides middleware-like transparent layer (mostly for programmers wishing to re-use functionality of the JBOWL package) for running of different tasks in a multi-threaded environment [5]. The available newly added services will be used e.g. for aspect-based sentiment analysis or formal concept analysis [4].

## 4 Application Cases

The aim of the application cases described below was to apply our framework into real practices to evaluate its usability and contributions to other state-of-the-art tools. Following four cases can be divided into two groups: the first two present recently finished experiments and the second pair sketches out our future research plans.

---

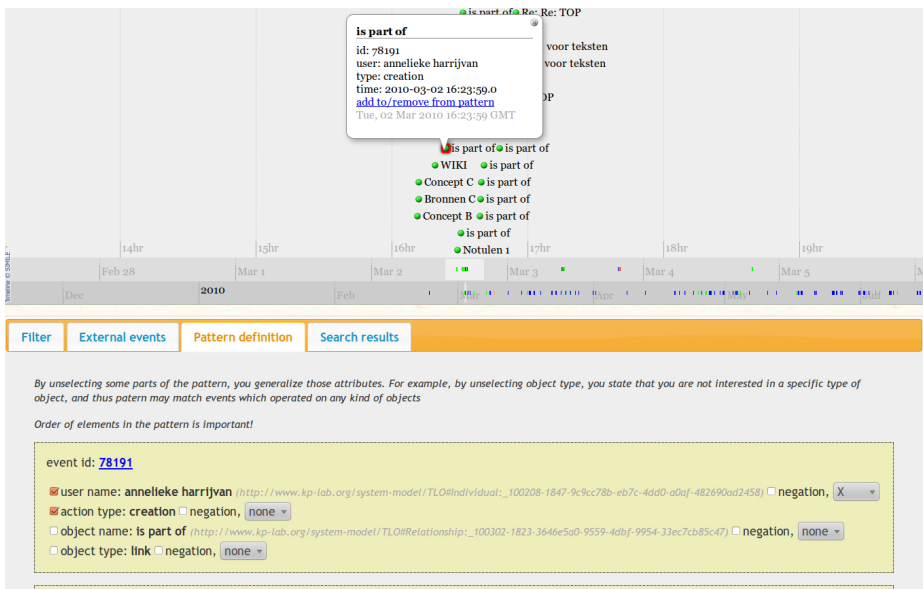
<sup>5</sup> <http://sourceforge.net/projects/jbowl>

## 4.1 KP-Lab System

KP-Lab System represents an initiative in the domain of collaborative learning environments and provides necessary functionalities to support knowledge building and creation within groups of users utilizing tools for collaborative editing, commenting, tagging, chatting, etc. This system was designed and implemented within European IST FP6 project called KP-Lab<sup>6</sup>. Our team participated on the development of logging and analytical services that were described in previous sections.

All performed actions in virtual user environment were monitored and stored into separate log repository in order to obtain necessary historical data for analytical purposes. At the end of the project (May 2011) log repository contained more than 100 thousand user activities. Even after official end of the project, KP-Lab System is still in operation and many of project's user partners use it in their courses.

The collected data can be used for two purposes: on-line notification services based on user specified requirements and off-line analyses in two different ways: quantitative analyses or timeline-based analyses with pattern discovery (see Fig. 2).



**Fig. 2.** Timeline-based visualization of one learning course within KP-Lab project with possibility to specify pattern definition

Experiments were based on real world data collected from pilot courses in Netherlands, Finland, Austria and Sweden. Analyses were performed by researchers and teachers of given courses, and they reflect genuine and real-world problems arisen in these courses. This included problems such as finding whether students

<sup>6</sup> <http://www.knowledgepractices.info/>

contribute to each others work, whether they take note of others findings and additions within a virtual shared space, or whether they cooperate on categorization of documents collected during the course. For each of the teacher's hypothesis, a formalized pattern was created and used in a subsequent search. For some problems, teachers were searching for occurrences of a given pattern, but sometimes even the number of results itself and frequency of result occurrences led to the resolution of the hypothesis. More detailed descriptions of performed experiments and their evaluation is provided e.g. in [12].

## 4.2 Wikipedia

Experiments were performed also on the data taken from the English Wikipedia. The sheer amount of logged information and contributing users and a collaborative nature of the Wikipedia provided a suitable environment for the evaluation of our analytical tools. The goal of our experiments was to analyze the cooperation of Wikipedia contributors in fighting the vandalism. We wanted to analyze and identify the methods for maintaining the quality of created knowledge.

Data for our analysis consisted from log of articles changes containing more than four hundred million events. Log contained changes of the 3.6 million of articles by 14 million users.

By textual analysis of event's free text description, we have identified events which have represented a fix of the vandalized article by undoing edits, also called reverting. Further preprocessing filtered out all other types of events and sorted the data in chronological order.

Traditional analytical tools applicable to Wikipedia offer only a minimal support for analysis across multiple users and articles. We have thus oriented on such kind of situations, and formulated following problems which we have than analyzed:

- Find a diligent user who reverted multiple articles vandalized by the same person.
- Find a persistent vandal, by finding situations where multiple people had to revert the actions of one vandal on one article.
- Identify vandals who defaced multiple articles over short period of time, in a way that multiple persons had to fix his edits.

We have formulated a formal pattern for each of the problems and analyzed the results of the search. In each case, results unanimously identified persons we were searching for, which we further confirmed by looking at users' profile, history of users actions or articles modified by given person.

The size of data was a challenge for this analysis. Moreover, events lacked the information about the type of action (whether it is a revert or a fix), and we had to analyze the comments attached to the events. Scalability of our analysis could be improved by incorporating a distributed storage and computation facilities into our tools, such as Hadoop and Cascalog.

### 4.3 Moodle

Moodle is one of the most used virtual learning systems in the world. Departments and faculties at our university use several Moodle portals for their learning courses. Even though Moodle offers some basic statistics about object usage frequency, it is not sufficient and we are in the process of integration with our framework in order to provide more advanced analyses. In this case, we performed some initial experiments with our logging services to obtain necessary historical data from Moodle. These experiments with the description of used methods and obtained results are described in [1]. We're planning to provide an effective combination of learning system and analytical extension, applicable for our teachers, researchers and students.

### 4.4 Software Testing Environment

Testing phase represents important step in the software development process and is typically supported by suitable technological solutions. Traditional testing environment offers functionalities to specify testing goals, testing procedures, input data, involved participants, expected goals, etc. The standard testing process contains sequence of planned user actions, but sometimes it is necessary to extend these sequences with measures to solve the unplanned or unexpected situations in software behavior. All performed actions and activities within testing process are monitored and stored in order to identify the advantages or disadvantages of used methods, approaches and ways of dealing with situation described before. This broad collection of historical data represents an interesting source for application of suitable analytical methods such as data mining, timeline-based visualization, pattern identification and discovery, etc. The successful realization of this case requires collaboration with companies engaged in software testing, in our case the RWE IT<sup>7</sup>. The cooperation has started with identification of possible application directions such as modeling of testing procedures and finding bugs based on the analysis of testing logs generated by automatic testing procedures.

## 5 Conclusion

In this paper we have examined a potential of analytical services for less structured collaborative activities realized within different virtual environments. The main aim of our proposal is to provide features for semi-automatic identification of important or interesting patters in collected historical data representing these activities. Its usability (sufficient content of the logs, effective logging mechanism, adequate pattern representation and its successful discovery) was tested within two initial sets of experiments: a collaborative learning system and Wikipedia. In the case of KP-Lab System, teachers and researches evaluated our framework as a very good step towards replacing manual examination of huge datasets after each learning course. Experiments with logs from Wikipedia were based on specified hypothesis that were

---

<sup>7</sup> <http://www.rweit-slovakia.com>

verified with our pattern services. Our future work will deal with adaptation of described framework into new conditions of software testing environment and we will also continue with research in a domain of learning systems, i.e. planned integration with Moodle and a dissemination within our university. The text mining services will be integrated in our analysis in order to acquire the additional information from produced text documents during realized activities.

**Acknowledgments.** The work presented in this paper was partially supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 (40%); the Slovak Research and Development Agency under the contract No. APVV-0208-10 (30%). This work is also the result of the project implementation Development of the Center of Information and Communication Technologies for Knowledge Systems (project number: 26220120030) supported by the Research & Development Operational Program funded by the ERDF (30%).

## References

1. Babič, F., Wagner, J., Jadlovská, S., Leško, P.: A logging mechanism for acquisition of real data from different collaborative systems for analytical purposes. In: SAMI 2010: 8th International Symposium on Applied Machine Intelligence and Informatics, Herľany, Slovakia, pp. 109–112. IEEE (2010)
2. Baker, R.S.J.D., Yacef, K.: The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining* 1(1), 3–17 (2009)
3. Bednár, P., Butka, P., Paralič, J.: Java Library for Support of Text Mining and Retrieval. In: Proceedings of the 4th Annual Conference Znalosti 2005, pp. 162–169 (2005)
4. Butka, P., Pócsová, J., Pócs, J.: A Proposal of the Information Retrieval System Based on the Generalized One-Sided Concept Lattices. In: Precup, R.-E., Kovács, S., Preitl, S., Petriu, E.M. (eds.) *Applied Computational Intelligence in Engineering*. TIEI, vol. 1, pp. 59–70. Springer, Heidelberg (2012)
5. Butka, P., Sarnovský, M., Bednár, P.: One Approach to Combination of FCA-based Local Conceptual Models for Text Analysis - Grid-based Approach. In: Proceedings of the 6th International Symposium on Applied Machine Intelligence, SAMI 2008, Herľany, Slovakia, pp. 131–135 (2008)
6. Harrer, A., Lingnau, A., Bientzle, M.: Interaction Analysis with dedicated logfile analysis tools – a comparative case using the PANdit tool versus manual inspection. In: Ninth IEEE International Conference on Advanced Learning Technologies, pp. 405–407. IEEE, Washington (2009)
7. Ilkhani, A., Abaee, G.: Extraction test cases by using data mining; reducing the cost of testing. In: Proc: Computer Information Systems and Industrial Management Applications CISIM 2010, Poland, pp. 620–625. IEEE (2010)
8. Martinez, A., Dimitriadis, Y., Rubia, B., Gomez, E., de la Fuente, P.: Combining qualitative evaluation and social network analysis for the study of classroom social interactions. *Computers and Education* 41(4), 353–368 (2003)
9. Muthyala, K., Naidu, R.: A novel approach to test suite reduction using data mining. *Indian Journal of Computer Science and Engineering* 2(3), 500–505 (2011)

10. Nurmela, K.A., Lehtinen, E., Palonen, T.: Evaluating CSCL log files by Social Network Analysis. In: Proc. CSCL 1999 Conference, pp. 434–444. Stanford University, Palo Alto (1999)
11. Paralič, J., Babič, F., Wagner, J., Simonenko, E., Spyrtatos, N., Sugibuchi, T.: Analyses of Knowledge Creation Processes Based on Different Types of Monitored Data. In: Rauch, J., Raš, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS, vol. 5722, pp. 321–330. Springer, Heidelberg (2009)
12. Paralič, J., Richter, C., Babič, F., Wagner, J., Raček, M.: Mirroring of knowledge practices based on user-defined patterns. *The Journal of Universal Computer Science* 17(10), 1474–1491 (2011)
13. Perera, D., Kay, J., Yacef, K., Koprinska, I., Zaiane, O.: Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering* 21(6), 759–772 (2009)
14. Rabbany, R., Takaffoli, M., Zaiāne, O.R.: Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. In: Proc. EDM, pp. 21–30 (2011)
15. Van der Aalst, W.M.P., et al.: ProM. The Process Mining Toolkit. In: Proceedings of the BPM 2009 Demonstration Track, Ulm, Germany. CEUR-WS.org, vol. 489 (2009)

# Lightweight Certificates

## – Towards a Practical Model for PKI\*

Łukasz Krzywiecki<sup>1</sup>, Przemysław Kubiak<sup>1</sup>, Mirosław Kutylowski<sup>1</sup>,  
Michał Tabor<sup>2,\*\*</sup>, and Daniel Wachnik<sup>2</sup>

<sup>1</sup> Faculty of Fundamental Problems of Technology, Wrocław University of Technology

<sup>2</sup> Trusted Information Consulting, Warsaw

{mirosław.kutylowski, lukasz.krzywiecki,  
przemysław.kubiak}@pwr.wroc.pl,  
{michal.tabor, daniel.wachnik}@ticons.pl

**Abstract.** We present a concept for Public Key Infrastructure based on certificates that are not understood as a guarantee of Certification Authority for unconditional authenticity of the data contained in the certificate. As liability of CA is a source of cost barrier for widespread use of PKI services, we concentrate on cost-efficient solutions. At the same time we formulate requirements that fill the security gaps of the traditional PKI. We present exemplary technical solutions that witness feasibility of these requirements.

**Keywords:** public key certificate, PKI, trust management, authentication, Schnorr signature.

## 1 Introduction

Asymmetric cryptography provides strong methods for proving that certain actions (like signing a document, remote authentication) are performed by a holder of a secret key related to a certain public key. However, for real life applications we have to link the holder of the secret key with a physical person. The purpose of PKI (Public Key Infrastructure) is to develop a framework, in which we can derive this linking information. The standard way is to issue digital certificates. Standards, such as X.509, help very much to develop software integrating cryptographic tools, however do not provide comparable security guarantees as a lot depends on purely procedural issues.

### 1.1 Traditional PKI Approach

The standard way of linking a physical person with a certificate is as follows:

- we assume that each person holds documents confirming identity, in most cases these are personal ID documents (personal ID documents issued by authorities, or less official ones like driving license, social security card, etc.). A silent assumption is that this system is secure.

---

\* This research has been partially supported by Foundation for Polish Science, Programme MISTRZ.

\*\* Corresponding author.

- There are Certification Authorities (CA) which issue certificates and which are obliged to fulfill many strict conditions.

**Actors** : person  $U$ , certification authority CA  
**Input** : a public key  $P$  of  $U$   
**Output** : certificate of the form “ $P$  is a public key of  $U$ ” signed by CA

```

1 begin
2    $U$  states a request for a certificate for key  $P$  the request is signed with the private key
   corresponding to  $P$ 
3   CA runs a procedure verifying that the request comes from  $U$ 
4   CA verifies the signature of the request with public key  $P$ 
5   if the outcome of verifications is positive, then CA issues a certificate for  $(P, U)$ 

```

**Algorithm 1.** Issuing a certificate - top level description

The process of issuing a certificate is described by Algorithm 1. The process mimics issuing documents like passports by public authorities. In most cases verification of the requester must be performed by personal appearance at a registration point of CA. One notable exception is the system of German personal identity cards, where this personal contact phase is overtaken by the authorities: the personal appearance occurs when the *Personalausweis* is given to the citizen. The protocols implemented on the ID card enable CA to check that the certificate request is coming from a personal ID document of the person declared in the request and that the correct password has been used. The fundamental assumptions for usage of such certificates are as follows:

- a person using a certificate can entrust the certificate and is not obliged to make any precaution steps (except for checking the current status of the certificate),
- if the certificate contains false information, CA is responsible for damages and liable to make compensations.

So far, the traditional PKI model failed as a business model despite all efforts of the European states to enforce widespread usage of digital signatures. In the next subsections we name some basic reasons for this situation.

**Economic Issues.** The running costs of CA as well as their business risk are to be compensated by the fees paid by the owners of the certificates. As issuing a certificate is a one-step service and there is no control over the actual usage of the certificate, **the only feasible payment policy is to charge per certificate**. In order to secure continuous income for CA (and for some security reasons as well), the certificates have limited validity period. The major problems for such a business model is that

a user has to pay a quite high entry price, while initially he may expect to use the certificate just a few times (at least in the first years)

There is a business deadlock: it makes no sense to buy a certificate if there are no services available, and there is no reason to build services, if the users are missing. Public administration may attempt to create such a market, but due to the cost issues usually the systems are downgraded to match justified requirements or they are not built at all.



**Security and Risk Management.** There are many risks of running CA services. Let us mention a few ones related to inspection of personal identity documents when a requester appears at an registration office:

- Getting a genuine personal identity document with false data is sometimes not that hard to get, as it is commonly believed.
- A genuine document with a modified photo might be used to get a certificate (at least as long as the identity documents are not equipped with electronic layer with authentication of biometric data). In this way a third person holding a private key may get possibility to sign documents in behalf of the person named in the certificate.
- There is a human factor risk – a verifier may err and oversee some issues that should stop issuing a certificate.
- There is no possibility to audit the verification process at a later time.

Last not least, cryptography used to sign certificates may become broken secretly, but as the concept is based on black-list approach, CA will even not know about forged certificates until there is a revocation attempt.

**Flexibility Issues.** One of the important features that are missing in the traditional PKI architecture is lack of flexibility. In particular:

- A certificate does not change during its validity period. This mimics the model of personal identity documents - so it must not contain any data of temporary validity.
- Any change concerning data confirmed in a certificate requires invalidating this certificate and issuing a new certificate. Such changes concern both the data about certificate owner as well as application policy on key's usage.
- The users need flexible ways of managing certificates - not only suspending a certificate in case of uncertainty of safety of signature creation device.
- There is a single control model - it is assumed that the owner keeps sole control over the signature creation device. This is legally confirmed by the status of the certificate, but far from reality, where gray zone situations dominate.
- All data contained in the signature are confirmed by CA. This leaves no room for entries of the type "*according to the declaration of the certificate owner*".
- There is no room for additional data supporting verification: verification is regarded as a sole problem of the trustee.

**Secure Signature Creation Device (SSCD).** In order to keep the link between a person and a private key, it is also necessary to assure that the person named in the certificate is the only person holding the private key. While it is quite hard to guarantee this as technical reality, there are attempts to solve the problem via appropriate legal assumptions. Again this leads to problems:

- CA issuing certificates may be obliged to present a list list of SSCDs. So in case of frauds, some part of responsibility can be assigned to CA. As a side effect, this increases the costs of certificates.

- The state runs a certification process for devices. In case of problems the users (signers and people entrusting the signed documents) are left alone with their problems, since suing a state is usually hopeless.
- In order to reduce costs, neither legislature nor CA's demand a signer to prove that he is actually holding a SSCD. The price to be paid is increasing risks of people entrusting electronic signatures. Since the most fundamental technical guarantees disappear, the real value of a signature and a certificate becomes questionable.

## 2 Lightweight Certificates - PKI 2.0 Approach

Below we report some ideas that we have developed within the initiative [1]. In order to build a well functioning PKI, we feel that the following fundamental properties must be fulfilled:

**No Entry Barrier:** The cost of receiving PKI services as a signatory must be correlated with the number of signatures created; there should be no initial fee covering all expenses of CA. However, in order to enable such an approach we have to rethink the technical part of system. In particular, we need a reliable accounting supported automatically by PKI.

**Multi-factor Security:** Rather than depend on a single security mechanism there should be multiple measures that are fairly independent and do not fail at the same time. The system should be composed of multiple components, so that the system should work even if some components fail.

**Explicit Security Declaration:** Security conditions should be explicitly declared, a signatory as well as trustee must have possibility to make their decisions concerning certain electronic signatures after estimating their personal risk.

**Transparency:** Security mechanisms should be based on principles that are available for evaluation to the public. It should be possible to test the strength of security mechanisms or detect irregularities during runtime of the system. Black-box solutions should be avoided; as a minimal condition they should be declared as such.

**No Trust Assumption:** No a-priori assumption should be made about trustfulness of any component of the system. Therefore, verification of the system must not be delegated to a third party that must be blindly trusted.

**Flexibility:** PKI should offer solutions for different cost and security levels, so that it may be adjusted to particular application case and risk analysis results. A user must be able to make decisions concerning services, authentication mechanisms and security level according to own needs and risks.

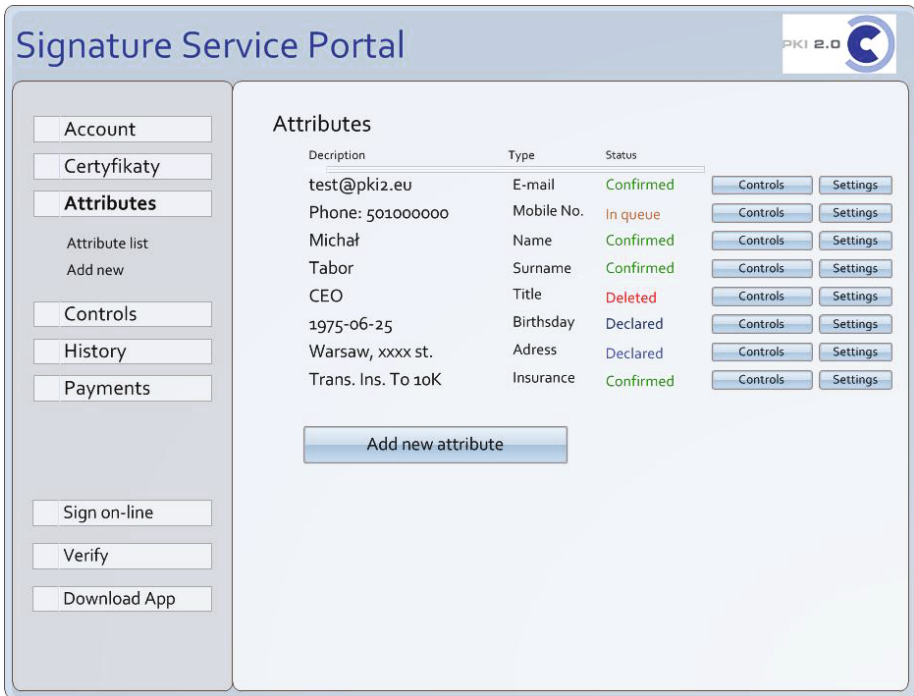


Fig. 1. User interface for certificate owner at CA

## 2.1 Lightweight Certification Process

In this section we give an overview of a system that aims to fulfill the conditions mentioned above (it depends also on technical components that are discussed in Sect. 3). The snapshot is given from the user's perspective. The items discussed below compose an exemplary menu of the user (see also Fig. 1):

**Account:** All activities concerning a user should occur within an account. Unlike in the traditional approach, creating an account should not be pre-conditioned by any requirements apart from simple mechanisms that prevent creating accounts by robots. Opening an account should follow the business model of accounts in such services as offered by Google. After opening an account, appropriate (lightweight) mechanisms should guarantee that:

- access to the account is given only to the owner,
- there are effective mechanisms to regain control over an hijacked account.

Login and password is a default, but it can be modified in *user authentication* section.

**Creating Certificates:** a user is free to request a certificate for any public key, conditioned by the proof the he really holds the private key. The certificate itself is not

automatically a guarantee that the person named in the certificate is the real holder of the private key (since there is no a priori authentication of identity). However, for each data field in question CA provides a statement about information's originator and status.

The most important concept is that a certificate can be modified at a later time when the status of the signatory changes. In some sense the changes only occur in the append mode: a certificate contain information about data confirmation, but always determines the time point of confirmation. As the situation may change over time, the certificate does not include information about the future – something like validity period of confirmation, as it is usually purely speculative.

**Attributes:** A user can freely create the list of his attributes that can be visible in his certificates. They can contain identification data such as email address, telephone number, web page, social networks account, as well as more formal one like personal ID number (if available), public keys for authentication with personal e-ID card, and so on. The main difference is that to each attribute we attach its status information. The status might have for instance the following form:

**declared:** the attribute has not been checked by CA and was included in the certificated based on declaration of the certificate owner.

**in queue:** confirmation of the certificate has been requested, but the process has not terminated yet,

**confirmed:** the attribute was verified positively by CA.

**User Authentication Controls:** In this section the user defines authentication methods. This includes not only access to the account (for more important operations), but also methods used for signature creation. Note that for some authentication methods CA is directly involved in the process.

Declaration of authentication methods in turn is a very important information for a person getting a document signed digitally. He can evaluate this information according to own policy and decide whether to entrust the document.

**Private Key Protection Mechanisms:** Since protection of secret keys is Achilles heel of digital signatures, it is fundamental to choose appropriate key protection mechanisms. Example choices are: cryptographic card (the traditional approach), storing keys on a secure server (as in Denmark), splitting the keys and storing key shares in different devices (see Sect. 3.2) or just storing them on a disk of PC (pgp approach).

**History:** One of basic methods of detecting irregularities is providing information concerning executed operations – as it is the case for bank accounts. Note that some of the methods described in Sect. 3 guarantee that no signature can be created without active participation of CA (or another third party as a service provider). Some of the signer's history might be available for proof purposes to people entrusting digital signatures.

**Payments:** Apart from its primary role, the list of payments has also some role in our trust model. So as before, a reduced information might be available to third parties.

### 3 Technical Tools - Enhancing Security with Schnorr Signatures

In this section we present a case study showing that it is feasible to implement the ideas presented in Sect. 2. Mostly, we build up our solutions mainly on top of Schnorr signatures - a scheme of higher flexibility than the standard ones like RSA and DSA family. As recently the patent of Schnorr signatures [2] has expired, there is a growing interest in this scheme ([3]). In Fig. 2 and 3 we recall this scheme.

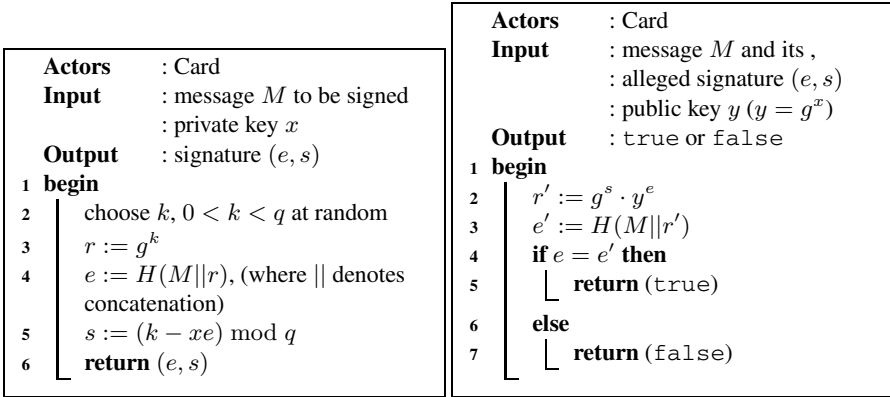


Fig. 2. Creating Schnorr signature with hash function  $H$

Fig. 3. Verification of a Schnorr signature

#### 3.1 Controlled Access

In the traditional approach an owner of SSCD cannot be certain about SSCD usage. In particular, if SSCD is a wireless smart card, then a malicious reader may attempt to activate it without knowledge of the owner. PIN protection only slows down the attacks. Algorithm 2 is a simple method that prevents signature creation by enforcing contact with a Monitoring Authority. A symmetric encryption scheme  $E$  must guarantee plaintext integrity.

#### 3.2 Mediated Signatures

The solution presented in Sect. 3.1 does not protect against leaking secret keys from SSCD. Mediated signatures go beyond this: the SSCD contains a private key, but there is another key outside SSCD that must be used in order to create a valid signature. Output of SSCD alone is worthless for a verification procedure.

Mediated signatures has been proposed originally for RSA [4]; a simple scheme based on Schnorr signatures has been proposed in [5]. Before we go into technical details we discuss applications of this idea:

- A signatory may use cryptographic card as SSCD, but also put an additional key on his PC. If the SSCD is kept separately from the PC, then probability of getting access to the private key is reduced significantly.

**Actors** : SSCD, Card Reader, Monitoring Authority, Signatory (optional)  
**Input** : message  $M$  to be signed  
: symmetric key  $w$  shared by SSCD and Monitoring Authority, SSCD holding serial number ID

**Output** : signature  $s$

**Procedure** : Executed by SSCD:

```

1 begin
2   optional: verify PIN
3   receive a request to sign  $M$ 
4   create permission request  $C := E_w(H(M), ID)$  for Monitoring Authority
5   send  $C$  to Reader
6   receive permission token  $C'$ 
7   if  $D(C') = \text{'yes, } H(M)\text{'}$  then
8     | create signature  $s$  of  $M$ , send  $s$  to Reader
9   else
10  | restart

```

**Procedure** : Executed by Card Reader:

```

11 begin
12  optional: send PIN to SSCD
13  send a request to sign  $M$ 
14  receive permission request  $C$ 
15  send  $C$  to Monitoring Authority
16  receive permission token  $C'$ 
17  send  $C'$  to SSCD
18  receive signature  $s$ 

```

**Procedure** : Executed by Monitoring Authority:

```

19 begin
20  receive permission request  $C$ 
21  decrypt  $C$  to  $H(M), ID$ 
22  check for ID on blacklists or white lists
23  optional: authenticate request with the signatory over an independent channel (like SMS) or inform the signatory about a signing attempt
24  if verification and/or check succeed then
25  |  $C' := E_w(\text{'yes, } H(M)\text{'})$ 
26  | send  $C'$  to Reader

```

**Algorithm 2.** Controlled signature creation

- Mediator scheme may be applied with Monitoring Authority as the second participant of the protocol. Then stealing the SSCD does not bring any advantage, if the Monitoring Authority is informed on time. Even if not, then some logs left by the Monitoring Authority may determine which signatures have been signed by a thief.
- Key generation of mediated Schnorr signature does not depend solely on neither Monitoring Authority nor on SSCD. So as long as they do not collude, mediated signatures are as safe as Schnorr signatures, even if one party is malicious.
- The scheme can be easily adopted to a  $k$ -party version. For instance, the secrets might be on a user's laptop, on a server in his office, and on a server of an external Monitoring Authority.

Now, let us recall the construction from [5]. The scheme uses a keyed hash function  $F$  with the range in  $\mathbb{Z}_q$ . First we describe key generation process.

<b>Actors</b>	: a signer $S$ , mediator $\mathcal{M}$ , key generation center KGC
<b>Input</b>	: a group $G$ of prime order $q$ for which DLP is hard, a generator $g$ of $G$ : signer's identity $S$ : a master secret $K$ of $\mathcal{M}$
<b>Output</b>	: private key $x$ for the signer $S$ , private key $x_S$ for $\mathcal{M}$ , public key $y$ for $S$
1	<b>begin</b>
2	$\mathcal{M}$ computes $x_S = F_K(S)$ and $y_S = g^S$
3	KGC chooses a number $x$ at random and computes $y_0 := g^x$
4	$S$ receives $y_S$ from $\mathcal{M}$ and $x, y_0$ from KGC; in both cases secure, authenticated channels are used
5	$S$ computes $y := y_0 \cdot y_S$ as its public key

**Algorithm 3.** Generating Keys for Mediated Schnorr Signatures

<b>Actors</b>	: a signer $S$ , mediator $\mathcal{M}$
<b>Input</b>	: a message $M$ to be signed : private key $x$ and public key $y$ of $S$ : private key $x_S$ of $\mathcal{M}$
<b>Output</b>	: signature $(e, s)$
1	$\mathcal{M}$ performs the following steps:
2	<b>begin</b>
3	choose $k_M, 0 < k_M < q$ , at random
4	$r_M := g^{k_M}$
5	compute a commitment $c = C(r_M)$ and sends it to $S$
6	$S$ performs the following steps:
7	<b>begin</b>
8	choose $k_S, 0 < k_S < q$ , at random
9	$r_S := g^{k_S}$
10	send $r_S$ together with the message $M$ (to be signed) to $\mathcal{M}$
11	$\mathcal{M}$ performs the following steps:
12	<b>begin</b>
13	$r := r_M \cdot r_S$
14	$e := H(M  r)$
15	$s_M := k_M - x_S \cdot e \bmod q$
16	send $r_M$ and $s_M$ to $S$
17	$S$ performs the following steps:
18	<b>begin</b>
19	check if $c = C(r_M)$
20	recompute $r := r_M \cdot r_S$ and $e := H(M  r)$
21	$s_S := k_S - x \cdot e \bmod q$
22	$s := s_S + s_M \bmod q$

**Algorithm 4.** Creating Mediated Schnorr Signature

Algorithm 4 executed by  $S$  and  $\mathcal{M}$  creates a regular Schnorr signature  $(e, s)$  given by equalities:  $r = g^{k_M + k_S}$ ,  $e = H(M||r)$  and  $s = (k_M + k_S) - (x + x_S) \cdot e \bmod q$ . They are correct since  $y = g^{x+x_S}$ .

In Algorithm 4 one can reduce the number of messages exchanged [5]. In this case security of the scheme is based on less standard assumption called Known-Target DLP. Otherwise, breaking the scheme is shown to lead to breaking Discrete Logarithm Problem in the random oracle model.

### 3.3 Separating Randomness and Key Usage

In the standard computation of Schnorr signature, computing  $s = k - ex \bmod q$  requires knowledge of secret key  $x$  and secret exponent  $k$ . If signature creation is implemented by one unit, then it is much easier to create hidden channels leaking the secret key.

<b>Actors</b>	: Randomization Unit, Key Unit
<b>Input</b>	: message $M$ to be signed : private key $x$ held by Key Unit, public key $y$ held by Randomization Unit
<b>Output</b>	: signature $(e, s)$
<b>Procedure</b>	: Executed by Randomization Unit:
1	<b>begin</b>
2	choose $k$ , $0 < k < q$ at random
3	$r := y^k$ , $e := H(M  r)$ , $z := k - e \bmod q$
4	send $z$ to Key Unit
5	return $e$ to signing application
	<b>Procedure</b> : Executed by Key Unit:
6	<b>begin</b>
7	receive $z$ from Randomization Unit
8	$s := x \cdot z \bmod q$
9	return $s$ to the signing application

**Algorithm 5.** An alternative way of creating Schnorr signatures

Schnorr signatures enable splitting the computation into two units, which we call Randomization Unit and Key Unit – see Algorithm 5. The first unit is responsible for deriving the component  $r = g^k$  to the signing application. The second one is responsible for operations with the secret key. So it is harder to create a hidden channel in a signature to leak the key by malicious hardware, and it is easier to certify an implementation. Any attack must be non-adaptive: an adversary must adapt  $k$  without seeing  $x$ .

### 3.4 Detecting Usage of Stolen Private Keys

The protocol presented below is a combination of ideas from [6] and [7]. The idea is that any attempt to create a signature with a leaked key leads to undeniable proof that the system has been compromised. Namely, if two Schnorr signatures use the same “randomness”  $k$ , then the private key  $x$  may be derived. To enable this feature, the scheme



uses commitments to predetermined random exponents that are used afterwards, during signature creation. We propose a procedure  $\text{Pre}(z)$  (Algorithm 7) that must be run by SSCD before the very first signature is created.  $\text{Pre}(z)$  outputs a list of commitments  $C$  (hashes  $\mathcal{H}(g^{k_i})$ , where  $\mathcal{H} : \{0, 1\} \rightarrow G$  is a strong hash function) that are going to be used for signature creation, exactly one commitment per signature. SSCD implements an internal counter which enables it to use each exponent/commitment only once. Thus if  $\text{Pre}(z)$  (see Algorithm 7) is run and commitments are published in some form (e.g. via a root of a Merkle tree), all subsequent verifiable signatures have to use it one after another. While creating the  $i$ th signature, SSCD executes Algorithm 2. The only difference is that SSCD uses the  $i$ th secret exponent  $k_i \in K$ , corresponding to the  $i$ th commitment  $h_i$ , instead of choosing  $k$  at random. Despite of exposure to chosen plaintext attacks, the scheme's security can be reduced to a standard assumption. The verification algorithm additionally checks whether the indicated commitment  $h_i$  corresponds to the exponent used. If an attacker that holds the secret key  $x$  of the user creates a signature of a message  $M'_i$ , according to some commitment  $i$ , the signature created by SSCD for the same commitment  $i$  can be used to prove forgery (see Algorithm 6).

<b>Input</b>	: Message $M$ and its alleged signature $(e, s, i)$ , Message $M'$ and its alleged signature $(e', s', i)$ , public key $y$ the commitment $h_i \in C$ from the sequence of predetermined exponents
<b>Output</b>	: true or false
1 <b>begin</b>	
2	$b := \text{Verify}(M, (e, s, i), y, C), \quad b' := \text{Verify}(M', (e', s', i), y, C)$
3	<b>if</b> $b$ and $b'$ and $(M \neq M')$ <b>then</b>
4	$x := (s' - s)/(e - e'),$ <b>return</b> $(x)$

**Algorithm 6.** Exposing secret  $x$  in case of forgery

<b>Actors</b>	: SSCD
<b>Input</b>	: The maximum number signatures $z$
<b>Output</b>	: A secret sequence $K$ of $z$ predetermined exponents, a public sequence $C$ of $z$ commitments to the predetermined exponents
1 <b>begin</b>	
2	<b>for</b> $i = 1, \dots, z$ <b>do</b>
3	$k_i \leftarrow_R \{1, \dots, q\}, \quad r_i := g^{k_i}, \quad h_i := \mathcal{H}(r_i)$
4	$K := \{k_i   i = 1, \dots, z\}, \quad C := \{h_i   i = 1, \dots, z\}$
5	<b>return</b> $K, C;$

**Algorithm 7.**  $\text{Pre}(z)$  – Exponent Predetermination

### 3.5 Pay Per Signature

*Pay per signature* approach can be implemented based on mechanisms from Sect. 3.1 and 3.2, as the Monitoring Authority *must* participate in signature creation and therefore knows the number of signatures created.

If the protocol has to create a proof verifiable by third parties, then we can slightly modify Algorithm 4 so that the Monitoring Authority computes the signature and sends it to the signatory. Similarly, Algorithm 2 can be changed so that the SSCD creates a signature, encrypts it with the public key of Monitoring Authority and sends the ciphertext to the reader. The only way to retrieve the signature is to submit it to Monitoring Authority. Such solutions have disadvantage that Monitoring Authority *knows* the signatures. However, for data protection reasons we may change the format of plaintext messages: instead of hashing  $M$  we may choose a string  $r$  at random and hash  $(M, r)$ .

## References

1. Initiative: Pki 2.0 (2011), <http://pki20.eu>
2. Schnorr, C.: Method for identifying subscribers and for generating and verifying electronic signatures in a data exchange system. U.S. Patent 4,995,082 (1991)
3. Bender, J., Dagdelen, Ö., Fischlin, M., Kügler, D.: The pace protocol for machine readable travel documents, and its security. In: Financial Cryptography 2012. LNCS. Springer (to appear, 2012)
4. Boneh, D., Ding, X., Tsudik, G., Wong, C.M.: Instantaneous revocation of security capabilities. In: USENIX Security Symposium (2001)
5. Nicolosi, A., Krohn, M.N., Dodis, Y., Mazières, D.: Proactive two-party signatures for user authentication. In: NDSS. The Internet Society (2003)
6. Błażkiewicz, P., Kubiak, P., Kutylowski, M.: Two-Head Dragon Protocol: Preventing Cloning of Signature Keys. In: Chen, L., Yung, M. (eds.) INTRUST 2010. LNCS, vol. 6802, pp. 173–188. Springer, Heidelberg (2011)
7. Choi, C.J., Kim, Z., Kim, K.: Schnorr signature scheme with restricted signing capability and its application. In: Computer Security Symposium (CSS), Kitakyushu, Japan, IPSJ, pp. 385–390 (2003)

# Control of Assistive Tools Using Voice Interface and Fuzzy Methods

Vytautas Rudzionis, Rytis Maskeliunas, and Tomas Rasymas

Department of Informatics, Vilnius University Kaunas Faculty, Lithuania  
Kaunas University of Technology, Lithuania  
vytautas.rudzionis@vukhf.lt, rytis.maskeliunas@ktu.lt

**Abstract.** The paper describes voice controlled multimodal assistive system with fuzzy control. Voice commands are often most convenient way to control various assistive tools. For full functionality voice commands need interpretation. The detection of voice boundaries in the long audio recording was implemented. The experimental results of fuzzy based indoor navigation system are presented in this article. The fuzzy control strategy presented bellow works on a given trajectory principle. The position of the device, the distance from the trajectory, orientation and control tasks are evaluated according to visual data. Paper presents the control model and algorithm of a real-life prototype.

**Keywords:** voice technology, voice command recognition, voice user interface, fuzzy methods, intelligent control.

## 1 Introduction

The speech is often the most natural, easiest to use and the most convenient way of human-machine interaction. Implicit richness of human speech communication gives the user many degrees of freedom for control and input of various devices. Voice user interfaces has a series of advantages for the control of various applications. The advantages of voice user interfaces were described and shown in various papers such as [1].

In recent months personal digital assistant Siri introduced by Apple as an integral part of iPhone4 series attracted a lot of attention. The essential property of Siri is user interface controlled by voice commands. The success of Siri is caused mainly by three components: improved speech recognizer, improved natural language modeler and semantic analyzer. All these three components enabled the implementation of voice user interface with the flexibility never seen before in practical applications available to the general public. It should be emphasized that speech recognition accuracy is the first and basic element of such types of systems since only good enough voice recognition will enable to implement semantic analyzer. Another very important characteristic of Siri is placed even in the name – personal digital assistant. It could be speculated that voice controlled user interfaces could be applied as a basis of personal assistive tools for the people with disabilities too.

Applications of speech technology can be grouped in the areas of access, control, communication and rehabilitation/therapy. For people with different impairments different types of speech technologies are more important: for people with visual impairments speech synthesis is essential as a way to access information, for people with hearing impairments perceptual speech processing and amplification are crucial, for other disabilities other areas of speech technology can be more important. But it is really difficult to find people with some sort of impairment that cannot benefit from one or another aspect of voice technology.

The main group of interest which needs is addressed in this study is the motor-handicapped people. The characteristic property of such category of people is that they often simply can't use traditional keypad based control systems independently or the use of such systems is significantly restricted. Environmental Control Systems (ECS) or Smart home control interfaces are available which address many elements of home management for disabled people, such as control of audio-visual equipment, telephones, household appliances, doors and curtains as well as the ability to summon assistance. Most ECSs utilize switch-scanning or keypad interfaces for control. More recently, ECSs with speech recognition have been introduced and a number of such systems are available on the market. Their success depends on a number of factors most important of them being maturity of voice processing technology used. Even better results could be achieved implementing multimodal approach – combining several different modalities to work in parallel or supplementing each other. In example, a multi-modal interaction framework using speech recognition and computer vision to model a new generation of interfaces in the residential environment was developed in [2]. The design is based on the use of simple visual clues and speech interaction. The latter system incorporates video information processing block which moves this system to the class of multimodal systems. Experience shows that motor-handicapped people are keen to use voice technology. This is especially true for people with hand movement restraints where the use of voice recognition is the only mode to transfer a computer control commands.

This paper presents a study about the possibilities to implement intelligent control methods for the one of the most important assistive tools used by the disabled persons – the wheelchair. The wheelchair is controlled using voice commands. The very important factor is the quality of the recognition. Paper deals with the accurate detection of the speech utterance boundaries during continuous audio input. Some of the voice commands need to be interpreted. E.g., such command as “turn to the right” requires find the necessary angle for the rotation. The angle depends on the location of various objects in the proximity of wheelchair. In such situations fuzzy based intelligent control methods are used to find the necessary decision in the scene. Chapter 2 of this paper describes the investigation of detection of speech boundaries in long audio recordings. Chapter 3 deals with the issues related with the implementation of fuzzy methods for wheelchair movements control.

## 2 Recognition of Voice Commands Using the Detection of Acoustic Events in the Long Utterances

One of the characteristic properties of voice user interfaces targeted to the disabled people controlling such assistive tools as the wheelchair or similar device is the necessity to operate in the continuous audio input mode. Most of the commercial tools implementing voice control (e.g. SIRI personal assistant manager) are using manually controlled audio interface: user needs to press button and then to speak (to say voice command) and often to release the button to indicate that utterance has ended and recognition should be started. In principle this approach may be implemented for the disabled people too but often it is unacceptable. People with some kind of disabilities such as motoric disabilities can't move hands easily or may have difficulties when pressing the necessary button or key. In this case we need to implement continuous audio mode: audio signal is recorded permanently and passed to the signal processing system for further analysis continuously.

Here we have two choices: first choice is to pass to the recognition system each part of the recorded audio signal and try to recognize acoustic-phonetic content of the recording (even to recognize the noise and some nonsense recordings) or to try to find those parts of audio recording where user of the assistive tool is speaking and to pass only those parts of the recording to the speech recognizer. The first approach is more likely to produce more recognition errors since the bigger variety of audio patterns needs to be recognized while the second one has the potential to be more accurate. But for the second approach to be accurate enough important issue to detect properly the places where the user of the system is speaking should be solved.

The detection of the boundaries of acoustic events such as utterances in the long recordings, utterances in the noisy environment or the phoneme boundaries within a word is one of the most fundamental problems in speech processing. A lot of activities were devoted to solve this problem. Various algorithms were proposed for the detection of speech and segmentation of spoken utterances, e.g. several methods could be found in [3-6]. Most of the algorithms exploit such spoken signal properties as the articulatory movement features or the differences between the actual signal spectrum and the spectrum prediction using its first or second order regression. Many methods also exploit signal energy changes as the factor. The selection of those features are based on the analysis of the physical properties of speech signal, e.g. articulatory movements features describe the particular structure of the speech signal spectrum which is typical only for the transitions between different acoustic events.

We proposed proprietary speech detection in noisy and long recordings method [7]. This algorithm proved to be accurate and robust enough for the segmentation of spoken speech for a wide range of SNRs and various classes of noise types. In [8] was used modified algorithm for the task of the acoustic event segmentation. The algorithm described here is further modification of the algorithm presented in previous study.

The speech is non-stationary process over longer time spans. The non-stationarity of speech signal is the result of different nature of different phonetic units. At the same time speech could be considered as a quasi-stationary process over shorter time periods (a time frame is shorter than 30 ms though it depends on the phonetic content of a signal). The basic idea of speech detection algorithms of is to find the places in the long recording where the statistical properties changes rapidly enough.

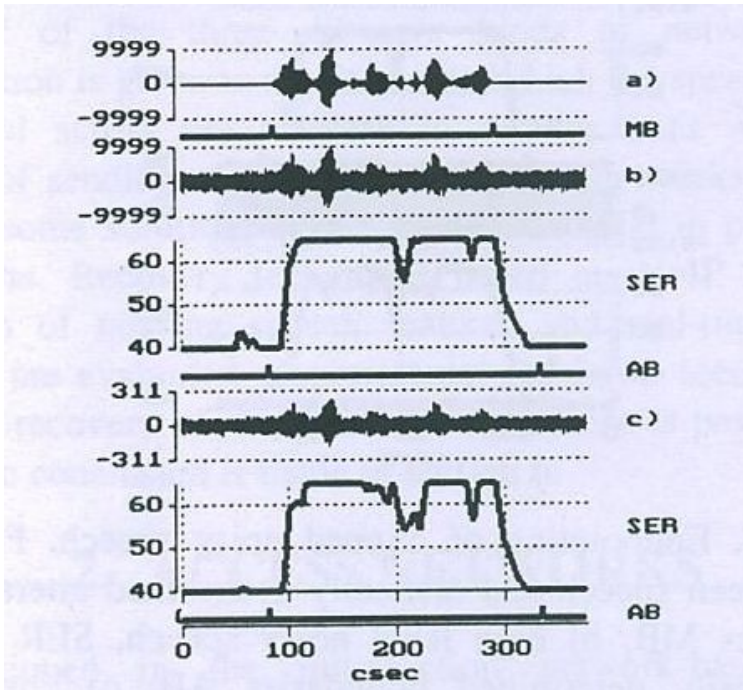
The algorithm for the detection of the acoustic events boundaries consists of several steps:

1. The logarithmic spectrum obtained from Fourier transform, linear prediction coding and IIR filter bank is derived using 8-10 ms step and 25 ms analysis window. If we will compare this algorithm with the algorithm used in earlier study we should note that the combination of three types of different spectrum was used. Logarithmic spectrum vectors were used to construct the likelihood function of the spectral rate changes in the utterance. The changes in the likelihood function values are used as the main indicator of speech presence at a particular time moment.
2. Then the spectral change rate function was filtered and integrated over experimentally set time spans. The integration allows obtain smoother likelihood function and helps avoid the random type of fluctuations which are characteristic for the changes in spectral properties of many phonetic units. The sequence of filtered and integrated parameters was used to detect the boundaries of acoustic events and was called an acoustic events response (AER): if its value exceeds experimentally defined threshold we will fix the start of the utterance and if the value it will drop below the threshold level for a experimentally set time period we will fix the end of utterance.

The Figure 1, which from top to down consists of the oscilogram of the original and the differentiated syllable, the spectrogram of the same syllable and the AER curve, is used to illustrate the efficiency of algorithm. It is expected that the changes in the acoustical content of a signal will occur on the places where the AER curve reaches the local maximum. The higher is the peak of the AER the higher is the likelihood value of the boundary between the different acoustic units.

We investigated the algorithm described above experimentally to find its efficiency in the voice commands recognition task. One group of experiments was carried on with a set of commands applicable for the control of wheelchair and using commercial speech recognition tools adapted to recognize Lithuanian voice command. Another group of experiments was carried on trying to define the limits of method. In this case a phonetically complicated material was used.

In the first group of experiments utterances of ten digit names (0-9) and a set of 24 voice commands that may be applied to control assistive device were used. The digit names were pronounced by 20 different speakers and each command was pronounced 20 times by each speaker. 24 control commands were read by 16 different speakers. Each speaker read every voice command 20 times as in the above case. Also the possibilities to adapt Microsoft commercial English and Spanish recognition engines were investigated in this context.



**Fig. 1.** Speech detection for various SNR. From top: clean speech and manually detected boundaries; low SNR signal, detected speech, higher SNR level, detected speech.

The voice command recognition experiments were carried on as follows. Long audio recordings were prepared (length of the recording is about 2 min). In each long recording voice command was inserted at a randomly selected place. The average length of voice commands was about one second and none of the tested commands was more than 2 sec in duration. To evaluate the impact of noise the additive white noise was added to the recording. In one case all recording was transferred to the recognizer and recognizer tried to recognize the acoustic content of whole recording. In the second case the whole recording was transferred to the utterance detection algorithm which tried to detect the part of the recording with the voice command present and only the detected voice command fragment was passed to the recognizer.

The Microsoft Speech API based commercial recognizers were used in this experiment. The English and Spanish recognition engines were used. For the adaptation to recognize Lithuanian voice commands methodology presented in [9] was used.

Table 1 shows the summarized results of this group of experiments. Here EN 7.0 means Microsoft English engine (ver 7.0) with ARPAbet-based phonetic transcriptions [10], ES 8.0 means that Microsoft Spanish engine (ver 8.0) with UPS-based transcriptions was used. Also Microsoft Speech Server English and Spanish engines (ver 9.0) (marked as EN9.0 and ES9.0 in the table) were used with UPS transcriptions.

**Table 1.** Recognition accuracy of voice commands in different environments and using different recognition engines (accuracy in percent)

Type of engine	Only voice command	Clean speech, SNR=30dB		Noisy speech, SNR=10 dB	
		Whole recording	Detected command	Whole recording	Detected command
EN7.0	87.4	79.3	85.6	58.6	78.8
ES8.0	94.6	88.3	92.3	66.5	85.6
EN9.0	77.0	74.5	76.8	51.2	63.4
ES9.0	97.0	75.6	97.0	80.4	92.3

The main conclusion which could be drawn from this table is that detection of voice commands helped to achieve higher overall recognition accuracy. The benefits of the voice command detection could be seen in all cases under investigation. Another observation is that the more complicated and noisy the acoustical environment is the more benefits could be get from the detection of voice command in the long utterance: if in the case with clean speech command detection allowed to achieve recognition accuracy increase by 10% in average then in the case with noisy speech increase in recognition accuracy was about 25% in average. Another observation is that better the base recognizer is (earlier studies showed that Spanish recognition engines are better suited for the adaptation to recognize Lithuanian voice commands than English engines; the reason is more similar to Lithuanian acoustic-phonetic structure of Spanish than English) the lesser the degradation of recognition accuracy in noisy environment could be expected.

Another group of experiments were performed using acoustically complicated and confusing data. These were utterances of syllables MA,NA,MI,NI,MO,NO known for their complicated phonetic structure and significant acoustic similarity to each other. Often such utterances are used to evaluate the potential of speech recognizer (so called boundary conditions). In these experiments phonetic material of 60 different speakers was used. Each speaker pronounced each syllable two times in isolation. One of those utterances was used for training (total 60 utterances training set) while another for testing (total 60 utterances testing set for each syllable). All utterances were manually processed marking the start and the end of acoustic phenomena. So the exact boundaries of each phonemic unit were known during training while for the testing they were known but weren't used.

In this case for the recognition was used CD-HMM based recognizer. The basic properties of HMM recognizer were straightforward: three states for each syllable and typical Baum-Welch reestimation procedure used to train the HMM model for each type of syllable. For the recognition also typical Viterbi search procedure was used. As for the acoustical front-end MFCC features were used to describe the acoustical structure of signal. MFCC features were supplemented with the change rate and acceleration coefficients (delta and delta-delta features) together with energy and energy delta. So the 39 features vector has been used. Output probabilities were modeled with single Gaussian and mixtures of several Gaussians.

Other methodology used in these experiments was the same as in the above described case: manually labeled clean recordings were used to get the reference



recognition accuracy; later utterances of syllables were inserted into the relatively long recordings at the random places. These recordings were transferred to the utterance detection algorithm and detected part then used for the recognition. Also whole utterances were used as above. In this case HMM model chains silence-syllable-silence or noise-syllable-noise was used. Table 2 summarizes the results obtained during this group of experiments.

**Table 2.** Recognition accuracy of nasalized syllables in different environments and using different recognition engines (accuracy in percent)

HMM model type	Manually labeled syllable	Clean speech, SNR=30dB		Noisy speech, SNR=10 dB	
		Whole recording	Detected command	Whole recording	Detected command
Single Gaussian	52.4	44.7	48.8	33.6	39.5
2 Gaussian mixtures	67.8	48.7	61.0	41.1	48.7
3 Gaussian mixtures	79.7	62.1	75.6	49.6	55.6
4 Gaussian mixtures	85.6	67.7	81.2	67.5	74.3
8 Gaussian mixtures	89.4	78.4	86.5	71.2	80.5

As could be seen from the results in Table 2 implementation of voice detection algorithm had significant impact to the recognition accuracy. Recognition gains were obtained for each syllable under consideration and both in the clean and noisy environments. The bigger gains were achieved in the presence of stronger noise and when simpler recognition algorithm was used (e.g. fewer Gaussian distributions were applied for the output probability modeling).

### 3 Fuzzy Methods for the Control of Assistive Tools

Control of assistive devices using voice commands is attractive solution. But only proper recognition of voice commands can't solve the needs of disabled people immediately in many cases. Let us imagine the control of movement of wheelchair by voice: the person in the wheelchair wants to make turn to the left. The key issue is how to select the desirable angle for the turn (it is obvious that the command "turn to the left" will not mean always turn by 90 degrees). It is possible to introduce several commands for making left turns or even to introduce the possibility to say exact angle in degrees by voice but all of these solutions aren't convenient in many cases: often it is difficult to determine the exact angle for the turn immediately and it is particularly inconvenient to correct inaccurate decisions later.

In our multimodal assistive device control platform we tried to introduce fuzzy methods using intelligent controller to find the appropriate (more exact) decision in particular environment. This decision depends on the environment and the position of other

objects in the nearby space. It means that when a command (let say “turn to the left”) has been recognized then simplified analysis of surrounding space is performed using video streams from cameras mounted on the head of the user and the distances to the objects are found. The intelligent controller based on fuzzy methods makes a decision what a turn angle should be. Below we will describe the fuzzy based intelligent control system used in multimodal platform to control wheelchair for disabled person.

It should be noted that a numerous attempts to implement intelligent methods for movement control of robotic devices were done in the past. Several examples of such attempts could be found in [11-14]. The fuzzy methods were selected because of their flexibility and good results achieved in various applications of similar kind and complicity as well as our experience working with them.

### 3.1 Control Model for the Assistive Platform

In this paragraph we present the kinematic model of the mobility platform illustrated in Figure 2.

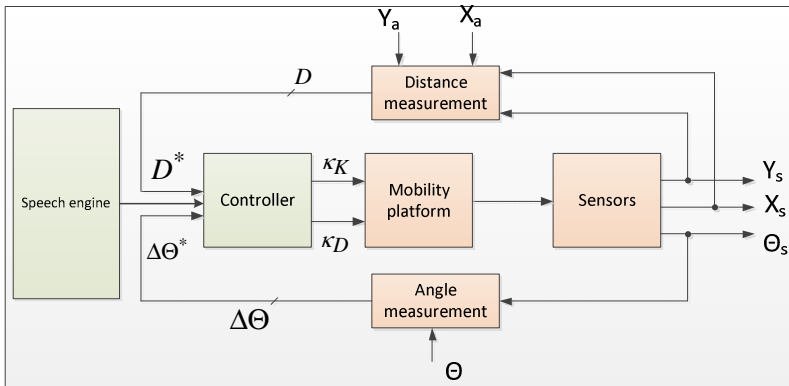


Fig. 2. Kinematic model of the wheelchair mobility platform

Think of it as a square with two sides capable of turning left and right by having a power vector on each side. The main movement rules mathematically could be written as follows.

$$\text{If } P_D = s \times K_{P_D} \text{ and } P_K = s \times K_{P_K} \text{ then } K_P = \frac{(P_D - P_K)}{I} \text{ and } P = \frac{(P_D + P_K)}{2}.$$

Here P – is a linear velocity, K – is the left side of the platform, D – is the right side of the platform,  $\kappa p$  – is the angular velocity, I – the length of an axle, s – the radius of a wheel. The dynamics of this system are described using the state space variables.

The movement of our mobility platform can be described like so  $K_D = \frac{1}{S}P + \frac{1}{2S}K$

and  $K_K = \frac{1}{S}P - \frac{1}{2S}K$ . The control of which is further modeled by voice.

### 3.2 Fuzzy Control Rules

The main purpose of the rules described below is to find the angle of rotation in depending on the location of objects in the environment used to move by the wheelchair. Naturally the system we are describing is non-linear, because of various factors such as time delay, noise, channel effects and other factors. The orientation  $\theta$  of the device is used as second input for the fuzzy controller and it shows how the device is oriented with respect to the trajectory segment. The orientation can get values from interval  $[-180^0, 180^0]$ . The mobility device is moving in good direction when orientation value is equal to 0. The device is moving towards trajectory when  $\theta$  is has a positive value, otherwise it moves from trajectory when  $\theta$  has a negative value. The angular velocity of the right power vector must be higher than the angular velocity of the left power vector, when the device is below or in the left and oriented away from the given trajectory. In the case, when the mobility device is above or in the right and oriented away from the given trajectory, the angular velocity of the left wheel must be higher than the angular velocity of the right wheel. The fuzzy logic is described in the control set shown Table 3.

**Table 3.** Fuzzy rules used for control

	$\Theta_H^-$	$\Theta_M^-$	$\Theta_Z$	$\Theta_M^+$	$\Theta_H^+$
$A_{\bar{H}}$	$\kappa_K^M$	$\kappa_K^H$	$\kappa_K^H$	$\kappa_K^H$	$\kappa_K^H$
	$\kappa_D^M$	$\kappa_D^M$	$\kappa_D^-$	$\kappa_D^-$	$\kappa_D^-$
$A_{\bar{M}}$	$\kappa_K^M$	$\kappa_K^M$	$\kappa_K^M$	$\kappa_K^H$	$\kappa_K^H$
	$\kappa_D^M$	$\kappa_D^Z$	$\kappa_D^Z$	$\kappa_D^Z$	$\kappa_D^-$
$A_Z$	$\kappa_K^-$	$\kappa_K^-$	$\kappa_K^M$	$\kappa_K^Z$	$\kappa_K^M$
	$\kappa_D^M$	$\kappa_D^Z$	$\kappa_D^M$	$\kappa_D^-$	$\kappa_D^-$
$A_M^+$	$\kappa_K^-$	$\kappa_K^Z$	$\kappa_K^Z$	$\kappa_K^Z$	$\kappa_K^M$
	$\kappa_D^H$	$\kappa_D^H$	$\kappa_D^M$	$\kappa_D^M$	$\kappa_D^-$
$A_H^+$	$\kappa_K^-$	$\kappa_K^-$	$\kappa_K^-$	$\kappa_K^M$	$\kappa_K^M$
	$\kappa_D^H$	$\kappa_D^H$	$\kappa_D^H$	$\kappa_D^H$	$\kappa_D^M$

The fuzzy controller has been designed with two inputs ( $A$  – the shortest distance of the center to the given trajectory,  $\Delta\Theta$  – angle between the trajectory and orientation line) and two outputs ( $\kappa_K$  - the angular velocity of the left side,  $\kappa_D$  - the angular velocity of the right side). The distance  $A$  is the first input described by 5 variables:  $A_{\bar{H}}$  .high negative,  $A_{\bar{M}}$  mean negative distances;  $D_Z$  - zero distance;  $A_M^+$  mean positive,  $A_H^+$  high positive distances. The angle  $\Delta\Theta$  is the second input similarly described by another 5 variables:  $\Delta\Theta_H^-$  .high negative,  $\Delta\Theta_M^-$  mean negative angles;  $\Delta\Theta_Z$  - zero

angle;  $\Delta\Theta_M^+$  average positive,  $\Delta\Theta_H^+$  high positive angles. Two outputs  $\kappa_K$  and  $\kappa_D$  are described by 3 variables:  $\kappa_{K,D}^Z$  - zero angular velocity of the wheels,  $\kappa_{K,D}^M$  - mean angular velocity of the wheels,  $\kappa_{K,D}^H$  - high angular velocity of the wheels.

## 4 Conclusions

Multimodal voice controlled assistive system for disabled people has been proposed. System allows recognize isolated commands together with some keywords. Important property of the proposed voice interface is that it works in permanent input mode and user did not need to press or release any key before speaking. To achieve this level of flexibility speech detection in long audio recordings algorithm was proposed. The algorithm is robust for a wide class of different noises and white range of SNRs. Experimental evaluation showed that using of speech detection algorithm allowed increase voice command recognition accuracy comparing with the recognition when detection wasn't used. Relative increase in the recognition accuracy was bigger in the case when the environment is noisy (noise type is additive and white). The increase of recognition accuracy was also observed in the acoustically complicated environments too. This observation allows us to make conclusion that recognition gains will be obtained using other sets of voice commands and will not be limited with the particular set implemented for the wheelchair control.

More convenient control of the assistive tools often requires better interpretation of recognized voice command. This is caused by the fact that some actions require the outcome which depends on the environment and other factors. For the wheelchair movement control intelligent control system was proposed.

The fuzzy logic based control strategy was deployed to execute the control, guiding an autonomous device along the indoor environments by using the robustness feature of the fuzzy controller design in the processing noisy and delayed data.

**Acknowledgments.** Parts of this work were done under the research project "Lietuviškų balso komandų atpažinimui orientuoto, multimodalinio išmaniųjų įrenginių asociatyvinio valdymo algoritmo sukūrimas ir modeliavimas", No.: 20101216-90 funded by EU SF project "Postdoctoral Fellowship Implementation in Lithuania" (VP1-3.1-ŠMM-01) of the Program of Human Resources Development Action Plan.

## References

1. Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Voice interactive systems. In: Helal, A., Mokhtari, M., Abdulrazak, B. (eds.) The Engineering Handbook of Smart Technology for Aging, Disability and Independence, pp. 281–297 (2008)
2. Macek, T., Kleindienst, J., Krchal, J., Seredi, L.: Multi-modal telephony services in home Intelligent Environments. In: 3rd IET International Conference, pp. 404–410 (2007)
3. Rabiner, L.R., Sambur, M.R.: An Algorithm For Determining the Endpoints in Isolated Utterances. Bell System Tech J. (54), 297–315 (1975)

4. Ying, G.S., Mitchell, C.D., Jamieson, L.: Endpoint Detection of Isolated Utterances Based on a Modified Teager Energy Measurement. In: Proc. of ICASSP 1993, pp. 732–735 (1993)
5. Hoyt, J., Wechsler, H.: Detection of Human Speech in Structured Noise. In: Proc. of ICASSP 1994, pp. 237–240 (1994)
6. Scheirer, E., Slaney, M.: Construction of Robust Multifeature Speech / Music Discriminator. In: Proc. of ICASSP 1997, pp. 1331–1334 (1997)
7. Rudzionis, A., Rudzionis, V.: Noisy speech detection and endpointing. In: Proc. of ISCA Workshop “Voice Operated Telecom Services”, Ghent, Belgium, pp. 79–84 (2000)
8. Rudžionis, V., Maskeliūnas, R., Rudžionis, A.: Assistive Tools for the Motor-Handicapped People Using Speech Technologies: Lithuanian Case. In: Abramowicz, W., Maciaszek, L., Węcel, K. (eds.) BIS 2011 Workshops, Part 2. LNBIP, vol. 97, pp. 123–131. Springer, Heidelberg (2011)
9. Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Investigation of foreign languages models for Lithuanian speech recognition. *Elektronika ir Elektrotechnika* 3, 15–20 (2009)
10. Jurafsky, D., Martin, J.: *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd edn. Prentice-Hall (2009)
11. Gokhan Ak, A., Cansever, G., Delibasi, A.: Robot Trajectory Tracking with Adaptive RBFNN-Based Fuzzy Sliding Mode Control. *Information Technology and Control* 40(2), 151–156 (2011)
12. Sun, D., Feng, G., Lam, C.M., Dong, H.: Orientation control of a differential mobile robot through wheel synchronization. *IEEE/ASME Trans. on Mechatronics* 10(3), 345–351 (2005)
13. Boquete, V., Garcia, R., Barea, R., Mazo, M.: Neural control of the movements of the wheelchair. *Journal of Intelligent and Robotic Systems*, 213–226 (1999)
14. Tang Shu-bo, Z., Yan, L., Lei, W.: Discrete trajectory tracking control of wheeled mobile robots. In: Proc. of the 2004 IEEE Int. Conf. on Robotics and Biometrics 2004, pp. 344–349 (2004)

# Author Index

- Ahmad, Iftikhar 236  
Alimazighi, Zaia 96
- Babič, František 284  
Barjis, Joseph 108  
Baumgrass, Anne 60  
Boukhedouma, Saida 96  
Buhl, Hans Ulrich 1
- Danenas, Paulius 249  
de Leoni, Massimiliano 48  
Di Ciccio, Claudio 11
- Eberhard, Robert 201
- Feldmann, Marius 189  
Fengel, Janina 36  
Fenz, Stefan 153
- Garsva, Gintautas 249  
Girit, Hasan 201  
Globa, Larysa 142
- Haller, Klaus 165  
Harrison-Broninski, Keith 120  
Heurix, Johannes 153  
Humm, Bernhard G. 36
- Iqbal, Javeria 236
- Janssen, Marijn 108
- Katz, Philipp 189  
Kleiner, Carsten 177  
Kochanowski, Monika 72  
Koetter, Falko 72  
Korhonen, Janne J. 120  
Koschel, Arne 177  
Kot, Tetiana 142  
Kowalkiewicz, Marek 132  
Krzywiecki, Lukasz 296  
Kubiak, Przemysław 296  
Kutyłowski, Mirosław 296
- Lehnert, Martin 1  
Leopold, Henrik 84
- Lisovskij, Karol 224  
Lunze, Torsten 189
- Maskeliunas, Rytis 308  
Matthes, Florian 165  
Mecella, Massimo 11  
Mending, Jan 84  
Michelberger, Bernd 201, 260  
Mutschler, Bela 201, 260
- Neubauer, Thomas 153
- Oussalah, Mourad 96
- Paralič, Ján 284  
Petrusel, Razvan 272  
Prabowo, Anisah Herdiyanti 108
- Raijers, Hajo A. 24  
Rasymas, Tomas 308  
Raudys, Aistis 224  
Reichert, Manfred 260  
Reverchuk, Andrey 142  
Rudzionis, Vytautas 308
- Schefer-Wenzl, Sigrid 60  
Schill, Alexander 142, 189  
Schmidt, Günter 236  
Schulz, Christopher 165  
Schunselaar, Dennis M.M. 24  
Sirvydis, Lukas 224  
Sprenger, Sebastian 189  
Stanciu, Paula Ligia 272  
Strasunskas, Darijus 213  
Strembeck, Mark 60
- Tabor, Michał 296  
Tamzalit, Dalila 96  
Tomassen, Stein L. 213
- Verbeek, Eric 24  
van der Aalst, Wil M.P. 24, 48  
van Dongen, Boudewijn F. 48
- Wachnik, Daniel 296  
Wagner, Jozef 284