

Mounia Lalmas Joemon Jose
Andreas Rauber Fabrizio Sebastiani
Ingo Frommholz (Eds.)

LNCS 6273

Research and Advanced Technology for Digital Libraries

14th European Conference, ECDL 2010
Glasgow, UK, September 2010
Proceedings

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Mounia Lalmas Joemon Jose
Andreas Rauber Fabrizio Sebastiani
Ingo Frommholz (Eds.)

Research and Advanced Technology for Digital Libraries

14th European Conference, ECDL 2010
Glasgow, UK, September 6-10, 2010
Proceedings

Volume Editors

Mounia Lalmas
Joemon Jose
Ingo Frommholz
University of Glasgow
Dept. of Computing Science
18 Lilybank Gardens, Glasgow, G12 8QQ, UK
E-mail: mounia@acm.org; {jj;ingo}@dcs.gla.ac.uk

Andreas Rauber
Vienna University of Technology
Dept. of Software Technology and Interactive Systems
Favoritenstr. 9-11, 1040 Vienna, Austria
E-mail: rauber@ifs.tuwien.ac.at

Fabrizio Sebastiani
Istituto di Scienza e Tecnologia dell'Informazione
Consiglio Nazionale delle Ricerche, Via G Moruzzi 1, 56124 Pisa, Italy
E-mail: Fabrizio.Sebastiani@isti.cnr.it

Library of Congress Control Number: 2010933080

CR Subject Classification (1998): H.4, H.2, H.3, H.5, J.1, H.2.8

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

ISSN 0302-9743
ISBN-10 3-642-15463-8 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-15463-8 Springer Berlin Heidelberg New York

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

springer.com

© Springer-Verlag Berlin Heidelberg 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

In the 14 years since its first edition back in 1997, the European Conference on Research and Advanced Technology for Digital Libraries (ECDL) has become the reference meeting for an interdisciplinary community of researchers and practitioners whose professional activities revolve around the theme of digital libraries. This volume contains the proceedings of ECDL 2010, the 14th conference in this series, which, following Pisa (1997), Heraklion (1998), Paris (1999), Lisbon (2000), Darmstadt (2001), Rome (2002), Trondheim (2003), Bath (2004), Vienna (2005), Alicante (2006), Budapest (2007), Aarhus (2008), and Corfu (2009), was held in Glasgow, UK, during September 6–10, 2010.

Aside from being the 14th edition of ECDL, this was also the last, at least with this name since starting with 2011, ECDL will be renamed (so as to avoid acronym conflicts with the European Computer Driving Licence) to TPLD, standing for the Conference on Theory and Practice of Digital Libraries. We hope you all will join us for TPDL 2011 in Berlin!

For ECDL 2010 separate calls for papers, posters and demos were issued, resulting in the submission to the conference of 102 full papers, 40 posters and 13 demos. This year, for the full papers, ECDL experimented with a novel, two-tier reviewing model, with the aim of further improving the quality of the resulting program. A first-tier Program Committee of 87 members was formed, and a further Senior Program Committee composed of 15 senior members of the DL community was set up. Each submitted paper was reviewed by four members of the first-tier PC, and a member of the Senior PC oversaw the process, stimulating discussion among the first-tier PC members in case of lack of consensus, providing her/his own “metareview” as well as a recommendation to the Program Chairs. All in all, each paper was thus carefully looked at by *five* experts, aside from the Program Chairs; we believe this resulted in a very accurate selection of the truly best submitted papers. Posters and demos were also evaluated by the same PC who evaluated papers, so as to increase the uniformity of evaluation standards. As a result, 22 long papers, 14 short papers, 19 posters and 9 demos were accepted, and are published in these proceedings. In addition, 14 submitted papers were accepted as posters.

The dense program of ECDL started on Monday with a range of tutorials providing in-depth coverage of both introductory as well as advanced topics in digital libraries. These included tutorials on the “Evaluation of Digital Libraries,” by Giannis Tsakonas and Christos Papatheodorou; on “Teaching/Learning About Digital Libraries,” by Edward Fox; on “Memento and Open Annotation,” by Michael L. Nelson, Robert Sanderson, and Herbert Van de Sompel; and on “Multimedia Document Access,” by Stefan R uger. On the same day, the Doctoral Consortium was held, where eight students presented their work and were given feedback by experts in digital libraries research.

The main conference featured keynote talks. One of them was given by Susan Dumais, from Microsoft Research, who explained how to understand and support people in interacting with dynamic information environments. Paper presentations were held in two parallel sessions, interleaved with the poster and demo sessions as well as one panel session on “Developing Services to Support Research Data Management and Sharing.” The panel members included Liz Lyon (moderator), Joy Davidson, Veerle Van den Eynden, Robin Rice, and Rob Grim.

Following the main conference, ECDL 2010 hosted three workshops, including the Workshop on Making Digital Libraries Interoperable (MDLI), the 9th Workshop on Networked Knowledge Organization Systems and Services (NKOS), and the Third Workshop on Very Large Digital Libraries (VLDL). For the first time in many years, the CLEF Workshop of the Cross-Language Evaluation Forum was not associated with ECDL, having spun off into a conference of its own, the CLEF Conference on Multilingual and Multimodal Information Access Evaluation (<http://clef2010.org/>) taking place just two weeks after ECDL 2010 in Padova, Italy. Best wishes to our “child conference” for a successful life on its own!

We would like to take the opportunity to thank everybody involved in making “the last ECDL” such an exciting event. Specifically, we would like to thank all conference participants and presenters, who provided a fascinating one-week program of high-quality presentations and intensive discussions, as well as all members from the Senior PC, the first-tier PC and the additional reviewers, who went to great lengths to ensure the high quality of this conference. Furthermore, we would like to thank all members of the Organizing Committee, and particularly everybody in the local organizing team at the University of Glasgow. Particularly, we would like to thank Keith van Rijsbergen who accepted to be our Honorary Chair, Matt Jones and Jaap Kamps who presided over the selection of posters and demos, Julio Gonzalo who dealt with the organization of the panel, Monica Landoni who dealt with the selection of tutorials, Jussi Karlgren who acted as Workshops Chair, Ian Anderson and Birger Larsen who were responsible for organizing a very interesting Doctoral Consortium, Maristella Agosti who (aside from providing guidance in her role as Chair of the ECDL Steering Committee) chaired the Best Paper Committee, Benjamin Piwowarski for his painstaking effort in compiling the proceedings, Vasiliki Kontaxi and Nick Duffield for their graphics work, and Tobias Blanke, Damaris Elsebach, Gabriella Kazai, Andrew McHugh, Seamus Ross and Ross Wilkinson, who—together with numerous student volunteers—assisted in various stages of organizing the conference. They all invested tremendous efforts to make sure that ECDL 2010 became an exciting and enjoyable event. The Conference and Visitor Services Office here in Glasgow were of great help in particular with respect to access to hotels and taking over the registration process. Finally, and not least, we are very grateful to our sponsors, DReSNet (EPSRC Digital Repositories e-Science Network), ExLibris, Yahoo! Research (sponsor of the best paper awards), CNI (Coalition for Networked Information), Glasgow City Marketing Bureau (who hosted the

welcome reception in the beautiful building of the Glasgow City Chambers), and the Humanities Advanced Technology & Information Institute (HATII) and the Department of Computing Science at the University of Glasgow (for hosting the ECDL website and their generous access to spaces and staff times). They allow us to keep costs down, which is very important in particular for students so that they can come, exchange ideas, learn and enjoy ECDL.

Finally, the General Chairs would like to thank Andreas Rauber and Fabrizio Sebastiani, the two Program Chairs, who worked very hard to ensure an excellent programme, and Ingo Frommholz, the Local Chair, for his dedicated contribution to the daily running and organization of the conference.

September 2010

Mounia Lalmas
Joemon Jose
Andreas Rauber
Fabrizio Sebastiani
Ingo Frommholz

Organization

ECDL 2010 was organized by the Department of Computing Science, University of Glasgow.

Organizing Committee

Honorary Chair

Keith van Rijsbergen University of Glasgow, UK

General Chairs

Joemon Jose University of Glasgow, UK

Mounia Lalmas University of Glasgow, UK

Local Chair

Ingo Frommholz University of Glasgow, UK

Program Chairs

Andreas Rauber Vienna University of Technology, Austria

Fabrizio Sebastiani Consiglio Nazionale delle Ricerche, Italy

Poster & Demo Chairs

Matt Jones University of Swansea, UK

Jaap Kamps University of Amsterdam, The Netherlands

Panel Chair

Julio Gonzalo UNED, Spain

Tutorial Chair

Monica Landoni University of Lugano, Switzerland

Workshop Chair

Jussi Karlgren Swedish Institute of Computer Science,
Sweden

Doctoral Consortium Chairs

Ian Anderson University of Glasgow, UK

Birger Larsen Royal School of Library and Information
Science, Denmark

Best Paper Award Chair

Maristella Agosti University of Padua, Italy

Proceedings Chair

Benjamin Piwowarski University of Glasgow, UK

Publicity Chairs

Gabriella Kazai Microsoft Research, UK
Andrew McHugh University of Glasgow, UK

Sponsor Chair

Tobias Blanke University of Glasgow and Kings College
London, UK

North and South America Liaison

Seamus Ross University of Toronto, Canada

Oceania and Asia Liaison

Ross Wilkinson Australian National Data Service, Australia

Local Organizing Committee

Damaris Elsebach Glasgow, UK

Graphics

Nick Duffield (Website Header)
Vasiliki Kontaxi (Logo)

Program Committee

Program Chairs

Andreas Rauber Vienna University of Technology, Austria
Fabrizio Sebastiani Consiglio Nazionale delle Ricerche, Italy

Senior Program Committee

David Bainbridge University of Waikato, New Zealand
George Buchanan City University of London, UK
Donatella Castelli Consiglio Nazionale delle Ricerche, Italy
Sally Jo Cunningham The University of Waikato, New Zealand
Edward Fox Virginia Polytechnic Institute and State
University, USA

Norbert Fuhr	University of Duisburg-Essen, Germany
Marcos Andre Goncalves	Federal University of Minas Gerais, Brazil
Stefan Gradmann	Humboldt University, Berlin, Germany
Carlo Meghini	Consiglio Nazionale delle Ricerche, Italy
Ray Larson	University of California, Berkeley, USA
Ingeborg Solvberg	Norwegian University of Technology and Science, Norway
Nicolas Spyratos	Université de Paris-Sud, France
Shigeo Sugimoto	University of Tsukuba, Japan
Hussein Suleman	University of Cape Town, South Africa
Elaine Toms	Dalhousie University, Canada

Programme Committees: Members and Reviewers

Trond Aalberg	Stefan Gradmann	Andras Micsik
Robert Allen	Jane Greenberg	Reagan Moore
Rodrigo Almeida	Preben Hansen	Atsuyuki Morishima
George Athanasopoulos	Donna Harman	Nektarios Moumoutzis
Joan Bartlett	Bernhard Haslhofer	Meinard Müller
Johan Bollen	Geneva Henry	Wolfgang Nejdl
José Borbinha	Jane Hunter	Michael Nelson
Christine Borgman	Antoine Isaac	Erich Neuhold
Pavel Braslavski	Min-Yen Kan	Robert Neumayer
Stephane Bressan	Noriko Kando	Liddy Nevile
Leonardo Candela	Maxi Kindling	David Nichols
Charles Cartledge	Ross King	Ragnar Nordlie
Ofelia Cervantes	Claus-Peter Klas	Kjetil Nrvg
Gobinda Chowdhury	Martin Klein	Vivien Petras
Sudatta Chowdhury	Traugott Koch	Ulrike Pfeil
Michael Christel	Dimitris Kotzinos	Dieter Pfoser
Stavros Christodoulakis	Laszlo Kovacs	Viet Phan-Luong
Vassilis Christophides	Alberto Laender	Andy Powell
Pierre Cubaud	Carl Lagoze	Edie Rasmussen
Theodore Dalamagas	Monica Landoni	Matthias Razum
Lois Delcambre	Audrey LaPlante	Harald Reiterer
Giorgio Maria Di Nunzio	Ronald Larsen	Laurent Romary
Susanne Dobratz	Dominique Laurent	Ian Ruthven
Stephen Downie	Ee-Peng Lim	Alfredo Sanchez
Peter Ecklund	Clifford Lynch	Robert Sanderson
Felix Engel	Akira Maeda	Felix Sasaki
Nicola Ferro	Thomas Mandl	Rudi Schmiede
Schubert Foo	Paolo Manghi	Timos Sellis
Richard Furuta	Hugo Manguinhas	Gianmaria Silvello
Giorgos Giannopoulos	Bruno Martins	Fabio Simeoni
Francois Goasdoue	Eva Mndez	Tim Smith

Ulrike Steffens	Yannis Tzitzikas	Megan Winget
Manfred Thaller	Vassilis Tzouvaras	Ian Witten
Yin-Leng Theng	Herbert Van de Sompel	Christian Wolff
Hellen Tibbo	Felisa Verdejo	Tim Wray
Anastasios Tombros	Jakob Voss	Paul Wu
Ricardo Torres	Paul Watry	Mohammad Zubair
Chrisa Tsinaraki	Barbara Wildemuth	

Workshops Program Committee

Carlos Castillo	Yahoo! Research, Barcelona, Spain
Nicola Ferro	University of Padua, Italy
Jussi Karlgren	Swedish Institute of Computer Science, Sweden (Chair)
Ian Soboroff	NIST, Gaithersburg, USA
Hanna Suominen	NICTA, Canberra, Australia

Tutorials Program Committee

Maristella Agosti	University of Padua, Italy
Monica Landoni	University of Lugano, Switzerland (Chair)
Ray R. Larson	University of California, Berkeley, USA

Doctoral Consortium Mentors

José Borbinha	Instituto Superior Técnico, Portugal
Laszlo Kovacs	Hungarian Academy of Sciences, Hungary
Milena Dobрева	University of Strathclyde, Glasgow, UK and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Giuseppina Vullo	University of Glasgow, UK

Panel Program Committee

George Buchanan	Centre for HCI Design, City University of London, UK
Julio Gonzalo	UNED, Spain (Chair)
Gregory Grefenstette	Exalead, Paris, France
Henning Müller	University Hospitals of Geneva, Switzerland
Nicola Orio	University of Padua, Italy

Sponsors



Table of Contents

Keynote: The Web Changes Everything: Understanding and Supporting People in Dynamic Information Environments	1
<i>Susan Dumais</i>	

System Architectures

Modelling Digital Libraries Based on Logic	2
<i>Carlo Meghini, Nicolas Spyratos, and Tsuyoshi Sugibuchi</i>	
General-Purpose Digital Library Content Laboratory Systems	14
<i>Paolo Manghi, Marko Mikulicic, Leonardo Candela, Michele Artini, and Alessia Bardi</i>	
Component-Based Authoring of Complex, Petri Net-Based Digital Library Infrastructure	22
<i>Yung Ah Park, Unmil Karadkar, and Richard Furuta</i>	

Metadata

Uncovering Hidden Qualities – Benefits of Quality Measures for Automatically Generated Metadata	30
<i>Sascha Tönnies and Wolf-Tilo Balke</i>	
Query Transformation in a CIDOC CRM Based Cultural Metadata Integration Environment	38
<i>Manolis Gergatsoulis, Lina Bountouri, Panorea Gaitanou, and Christos Papatheodorou</i>	
User-Contributed Descriptive Metadata for Libraries and Cultural Institutions	46
<i>Michael A. Zarro and Robert B. Allen</i>	

Multimedia IR

An Approach to Content-Based Image Retrieval Based on the Lucene Search Engine Library	55
<i>Claudio Gennaro, Giuseppe Amato, Paolo Bolettieri, and Pasquale Savino</i>	
Evaluation Constructs for Visual Video Summaries	67
<i>Stina Westman</i>	

Visual Expression for Organizing and Accessing Music Collections in MusicWiz	80
<i>Konstantinos Meintanis and Frank M. Shipman</i>	

Interaction and Interoperability

An Architecture for Supporting RFID-Enhanced Interactions in Digital Libraries	92
<i>George Buchanan and Jennifer Pearson</i>	
New Evidence on the Interoperability of Information Systems within UK Universities	104
<i>Kathleen Menzies, Duncan Birrell, and Gordon Dunsire</i>	
Enhancing Digital Libraries with Social Navigation: The Case of Ensemble	116
<i>Peter Brusilovsky, Lillian Cassel, Lois Delcambre, Edward Fox, Richard Furuta, Daniel D. Garcia, Frank M. Shipman III, Paul Bogen, and Michael Yudelson</i>	

Digital Preservation

Automating Logical Preservation for Small Institutions with Hoppla	124
<i>Stephan Strodl, Petar Petrov, Michael Greifeneder, and Andreas Rauber</i>	
Estimating Digitization Costs in Digital Libraries Using DiCoMo	136
<i>Alejandro Bia, Rafael Muñoz, and Jaime Gómez</i>	
In Pursuit of an Expressive Vocabulary for Preserved New Media Art	148
<i>Andrew McHugh and Leo Konstantelos</i>	

Social Web/Web 2.0

Privacy-Aware Folksonomies	156
<i>Clemens Heidinger, Erik Buchmann, Matthias Huber, Klemens Böhm, and Jörn Müller-Quade</i>	
Seamless Web Editing for Curated Content	168
<i>David Bainbridge and Brook J. Novak</i>	
Automatic Classification of Social Tags	176
<i>Christian Wartena</i>	

Search in Digital Libraries

Exploring the Impact of Search Interface Features on Search Tasks	184
<i>Abdigani Diriye, Ann Blandford, and Anastasios Tombros</i>	
Relevance in Technicolor	196
<i>Ulises Cerviño Beresi, Yunhyong Kim, Dawei Song, Ian Ruthven, and Mark Baillie</i>	
Application of Session Analysis to Search Interface Design	208
<i>Cathal Hoare and Humphrey Sorensen</i>	

(Meta) Analysis of Digital Libraries

An Analysis of the Evolving Coverage of Computer Science Sub-fields in the DBLP Digital Library	216
<i>Florian Reitz and Oliver Hoffmann</i>	
Analysis of Computer Science Communities Based on DBLP	228
<i>Maria Biryukov and Cailing Dong</i>	
Citation Graph Based Ranking in Invenio	236
<i>Ludmila Marian, Jean-Yves Le Meur, Martin Rajman, and Martin Vesely</i>	

Query Log Analysis

A Search Log-Based Approach to Evaluation	248
<i>Junte Zhang and Jaap Kamps</i>	
Determining Time of Queries for Re-ranking Search Results	261
<i>Nattiya Kanhabua and Kjetil Nørvgå</i>	
Ranking Entities Using Web Search Query Logs	273
<i>Bodo Billerbeck, Gianluca Demartini, Claudiu S. Firan, Tereza Iofciu, and Ralf Krestel</i>	

Cooperative Work in DLs

Examining Group Work: Implications for the Digital Library as Sharium	282
<i>Sandra Toze and Elaine G. Toms</i>	
Architecture for a Collaborative Research Environment Based on Reading List Sharing	294
<i>Gabriella Kazai, Paolo Manghi, Katerina Iatropoulou, Tim Haughton, Marko Mikulicic, Antonis Lempesis, Natasa Milic-Frayling, and Natalia Manola</i>	

CritSpace: A Workspace for Critical Engagement within Cultural
Heritage Digital Libraries 307
Neal Audenaert, George Lucchese, and Richard Furuta

Ontologies

German Encyclopedia Alignment Based on Information Retrieval
Techniques 315
Roman Kern and Michael Granitzer

Lightweight Parsing of Classifications into Lightweight Ontologies 327
Aliaksandr Autayeu, Fausto Giunchiglia, and Pierre Andrews

Measuring Effectiveness of Geographic IR Systems in Digital
Libraries: Evaluation Framework and Case Study 340
*Damien Palacio, Guillaume Cabanac, Christian Sallaberry, and
Gilles Hubert*

Domain-Specific DLs

A Visual Digital Library Approach for Time-Oriented Scientific
Primary Data 352
*Jürgen Bernard, Jan Brase, Dieter Fellner, Oliver Koepler,
Jörn Kohlhammer, Tobias Ruppert, Tobias Schreck, and Irina Sens*

DINAH, a Philological Platform for the Construction of
Multi-structured Documents 364
Pierre-Édouard Portier and Sylvie Calabretto

The PROBADO Project - Approach and Lessons Learned in Building
a Digital Library System for Heterogeneous Non-textual Documents 376
*René Berndt, Ina Blümel, Michael Clausen, David Damm,
Jürgen Diet, Dieter Fellner, Christian Fremerey, Reinhard Klein,
Frank Krahl, Maximilian Scherer, Tobias Schreck, Irina Sens,
Verena Thomas, and Raoul Wessel*

Posters

Capacity-Constrained Query Formulation 384
Matthias Hagen and Benno Maria Stein

AAT-Taiwan: Toward a Multilingual Access to Cultural Objects 389
Shu-Jiun Chen, Diane Wu, Pei-Wen Peng, and Yung-Ting Chang

Using Pattern Language as a Framework for Future Metadata
Structure 393
Esben Agerbæk Black

i-TEL-u: A Query Suggestion Tool for Integrating Heterogeneous Contexts in a Digital Library	397
<i>Maristella Agosti, Davide Cisco, Giorgio Maria Di Nunzio, Ivano Masiero, and Massimo Melucci</i>	
The Planets Testbed: A Collaborative Research Environment for Digital Preservation	401
<i>Brian Aitken, Seamus Ross, Andrew Lindley, Edith Michaeler, Andrew Jackson, and Maurice van den Dobbelsteen</i>	
A Functionality Perspective on Digital Library Interoperability	405
<i>George Athanassopoulos, Edward Fox, Yannis Ioannidis, George Kakalettris, Natalia Manola, Carlo Meghini, Andreas Rauber, and Dagobert Soergel</i>	
Overview and Results of the INEX 2009 Interactive Track	409
<i>Thomas Beckers, Norbert Fuhr, Nils Pharo, Ragnar Nordlie, and Khairun Nisa Fachry</i>	
SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size)	413
<i>Jöran Beel, Bela Gipp, Ammar Shaker, and Nick Friedrich</i>	
Academic Publication Management with PUMA – Collect, Organize and Share Publications	417
<i>Dominik Benz, Andreas Hotho, Robert Jäschke, Gerd Stumme, Axel Halle, Angela Gerlach Sanches Lima, Helge Steenweg, and Sven Stefani</i>	
Using Mind Maps to Model Semistructured Documents	421
<i>Alejandro Bia, Rafael Muñoz, and Jaime Gómez</i>	
Towards a Public Library Digital Service Taxonomy	425
<i>Steven Buchanan and David McMenemy</i>	
Multimodal Image Collection Visualization Using Non-negative Matrix Factorization	429
<i>Jorge E. Camargo, Juan C. Caicedo, and Fabio A. González</i>	
A New Perspective on Collection Selection	433
<i>Helen Dodd, George Buchanan, and Matt Jones</i>	
Creating a Flexible Preservation Infrastructure for Electronic Records	437
<i>Karen Estlund and Heather Briston</i>	
Matching Intellectual Works for Rights Management in the European Library	441
<i>Nuno Freire</i>	

Mopseus – A Digital Library Management System Focused on Preservation	445
<i>Dimitris Gavrilis, Christos Papatheodorou, Panos Constantopoulos, and Stavros Angelis</i>	
Link Proximity Analysis - Clustering Websites by Examining Link Proximity	449
<i>Bela Gipp, Adriana Taylor, and Jöran Beel</i>	
SliDL: A Slide Digital Library Supporting Content Reuse in Presentations	453
<i>José H. Canós, María Isabel Marante, and Manuel Llavador</i>	
Metadata Impact on Research Paper Similarity	457
<i>Germán Hurtado Martín, Steven Schockaert, Chris Cornelis, and Helga Naessens</i>	
Exploring the Influence of Tagging Motivation on Tagging Behavior	461
<i>Roman Kern, Christian Korner, and Markus Strohmaier</i>	
A Teaching Tool for Parasitology: Enhancing Learning with Annotation and Image Retrieval	466
<i>Nádia P. Kozievitch, Ricardo da Silva Torres, Felipe Andrade, Uma Murthy, Edward Fox, and Eric Hallerman</i>	
Framework for Logging and Exploiting the Information Retrieval Dialog	470
<i>Paul Landwich, Claus-Peter Klas, and Matthias Hemmje</i>	
Defining the Dynamicity and Diversity of Text Collections	474
<i>Ilya Markov and Fabio Crestani</i>	
Manuzio: A Model for Digital Annotated Text and Its Query/Programming Language	478
<i>Marek Maurizio and Renzo Orsini</i>	
Effective Term Weighting for Sentence Retrieval	482
<i>Saeedeh Momtazi, Matthew Lease, and Dietrich Klakow</i>	
User-Oriented Evaluation of Color Descriptors for Web Image Retrieval	486
<i>Otávio A.B. Penatti and Ricardo da S. Torres</i>	
A Topic-Specific Web Search System Focusing on Quality Pages	490
<i>Ari Pirkola and Tuomas Talvensaari</i>	
Reliable Preservation of Interactive Environments and Workflows	494
<i>Klaus Rechert, Dirk von Suchodoletz, Randolph Welte, Felix Ruzzoli, and Isgandar Valizada</i>	

Automated Country Name Disambiguation for Code Set Alignment	498
<i>Gramm Richardson</i>	
LIFE-SHARE Project: Developing a Digitisation Strategy Toolkit	502
<i>Beccy Shipman, Matthew Herring, Ned Potter, and Bo Middleton</i>	
Ensemble: A Distributed Portal for the Distributed Community of Computing Education	506
<i>Frank M. Shipman, Lillian Cassel, Edward Fox, Richard Furuta, Lois Delcambre, Peter Brusilovsky, B. Stephen Carpenter II, Gregory Hislop, Stephen Edwards, and Daniel D. Garcia</i>	
A New Focus on End Users: Eye-Tracking Analysis for Digital Libraries	510
<i>Jonathan Sykes, Milena Dobрева, Duncan Birrell, Emma McCulloch, Ian Ruthven, Yurdagül Ünal, and Pierluigi Feliciati</i>	
Digital Library Educational Module Development Strategies and Sustainable Enhancement by the Community	514
<i>Seungwon Yang, Tarek Kanan, and Edward Fox</i>	
Demos	
Approach to Cross-Language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database	518
<i>Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda</i>	
Open Source Historical OCR: The OCRopodium Project	522
<i>Michael Bryant, Tobias Blanke, Mark Hedges, and Richard Palmer</i>	
A Voice-Oriented Image Cataloguing Environment	526
<i>José H. Canós, Carlos J. Castillo, Pablo Muñoz, Héctor Valero, and Manuel Llavador</i>	
DMP Online: A Demonstration of the Digital Curation Centre's Web-Based Tool for Creating, Maintaining and Exporting Data Management Plans	530
<i>Martin Donnelly, Sarah Jones, and John W. Pattenden-Fail</i>	
DiLiA – The Digital Library Assistant	534
<i>Kathrin Eichler, Holmer Hensen, Günter Neumann, Norbert Reithinger, Sven Schmeier, Kinga Schumacher, and Inessa Seifert</i>	
Xeproc©: A Model-Based Approach towards Document Process Preservation	538
<i>Thierry Jacquin, Hervé Déjean, and Jean-Pierre Chanod</i>	

A Prototype Personalization System for the European Library Portal . . .	542
<i>Marialena Kyriakidi, Lefteris Stamatogiannakis, Mei Li Triantafyllidi, Maria Vayanou, and Yannis Ioannidis</i>	
Meta-Composer: Synthesizing Online FRBR Works from Library Resources	546
<i>Michalis Sfakakis, Panagiotis Staikos, and Sarantos Kapidakis</i>	
Digital Library in a 3D Virtual World: The Digital Bleek and Lloyd Collection in Second Life	550
<i>Rizmari Versfeld, Spencer Lee, Edward Fox, Hussein Suleman, and Kyle Williams</i>	
Appendix:	
Doctoral Consortium, Workshops, Tutorials, and Panel	555
Author Index	571

Keynote: The Web Changes Everything: Understanding and Supporting People in Dynamic Information Environments

Susan Dumais

Microsoft Research
sdumais@microsoft.com

<http://research.microsoft.com/~sdumais>

Abstract. Most digital library resources and the Web more generally are dynamic and ever-changing collections of information. However, most of the tools that have been developed for interacting with Web and DL content, such as browsers and search engines, focus on a single static snapshot of the information. In this talk, I will present analyses of how web content changes over time, how people re-visit web pages over time, and how re-visitation patterns are influenced by user intent and changes in content. These results have implications for many aspects of search including crawling, ranking algorithms, result presentation and evaluation. I will describe a prototype that supports people in understanding how information they interact with changes over time, by highlighting what content has changed since their last visit. Finally, I will describe a new retrieval model that represents features about the temporal evolution of content to inform crawl policy and improve ranking.

Modelling Digital Libraries Based on Logic

Carlo Meghini¹, Nicolas Spyratos², and Tsuyoshi Sugibuchi²

¹ Consiglio Nazionale delle Ricerche, Istituto della Scienza e delle Tecnologie della Informazione, Pisa, Italy

² Université Paris-Sud, Laboratoire de Recherche en Informatique, Orsay Cedex, France

Abstract. We present a data model for digital libraries supporting identification, description and discovery of digital objects. The model is formalized as a first-order theory, certain models of which correspond to the intuitive notion of digital library. Our main objective is to lay the foundations for the design of an API offering the above functionality. Additionally, we use our formal framework to discuss the adequacy of the Resource Description Framework with respect to the requirements of digital libraries.

1 Introduction

Today, digital information comes in the form of digital objects such as JPEG images or PDF documents, and these objects can be assembled into more complex objects. For example, a photograph in digital form, accompanied by a caption (also in digital form) is a complex object consisting of two other objects, the picture and the caption.

In general, a complex object is created from simpler objects that collectively form a meaningful unit. The simpler objects might have been created from scratch or they may belong to other complex objects and are simply extracted from or referenced within these other complex objects and re-used in order to create the new complex object. For example, suppose that one has a collection of photos from last summer's vacations, each with an accompanying caption, and wants to create a digital album including the best of these pictures. The so created album, together with a caption such as "Summer 2009", would be a new digital object composed of other digital objects. To do so, one needs to be able to access the individual objects in order to select the pictures, and then associate the selected pictures collectively with a newly created caption, thus forming a new complex object.

Roughly speaking, what we call a digital library (DL for short) is a set of digital objects and services that allow a community of users to access and re-use the objects. In particular, each object in the DL is associated with a *content* and a number of *descriptions*. The users of the DL should be able to perform the following tasks:

- *describe* an object of interest according to some vocabulary;
- *discover* objects of interest based on content and/or description;
- *identify* an object of interest, in the sense of assigning to it an identity;
- *re-use* objects in a different context (*e.g.* by adding them to the content of existing objects or by creating new complex objects).

We note that the concept of DL constitutes a major departure from that of a traditional information system. Indeed, a traditional information system contains *representations* of real-world objects (such as employees and departments), while a DL might contain both, representations of objects and the objects themselves.

In this paper, we present a data model for DLs supporting identification, description and discovery of digital objects. The model is formalized as a first-order theory, certain models of which correspond to the intuitive notion of DL. Our main objective is to lay the foundations for the design of an API offering the above functionality. The implementation of the API will form the core of a DL management system [3]. Similarly to a database management system, a DL management system should relieve application developers from the burdens of implementing the basic functionality of a DL so that they can focus on the design and implementation of applications.

Additionally, we use our formal framework to discuss the adequacy of the Resource Description Framework with respect to the requirements of DLs.

The rest of the paper is structured as follows: Section 2 provides an informal definition of a DL, while Section 3 presents the formal model along with an example. Section 4 presents the query language and Section 5.1 discusses RDF. Section 6 briefly reviews some relevant literature, and finally, Section 7 concludes the paper and outlines future work. Preliminary versions of parts of this work have been presented at various events [11,12,13].

2 Digital Libraries: An Informal Definition

The basic notion of our model is that of a *digital object* (or simply *object*). Intuitively, we think of a digital object as a piece of information in digital form such as a text, an executable piece of software, a URI and so on. As such, a digital object can be processed by a computer, for instance it can be stored in memory and displayed on a screen. For the purposes of our discussions, we assume the existence of a (countable) set consisting of all digital objects that one can ever define. We shall denote this set as \mathcal{O} .

We view a DL as a finite subset of \mathcal{O} , in which each object o is associated with a *content* and a set of *descriptions*. The *content* of an object o is the set of objects which make up o . For example, the content of a book is the set of its chapters, each chapter being an individual object. Similarly, the content of an exhibition of paintings is the set of paintings in the exhibition. A *description* of an object o is an assignment of values to some characteristics of the object. For example, the description of a book will include the title and the author of the book, while the description of a painting exhibition will include the date and the place of the exhibition. A book, however, can be described from different points of view, each leading to a different description. As a result, a book might be associated to a set of descriptions (and the same goes for a painting exhibition). In order to accommodate several descriptions for the same object, we will treat descriptions as objects in their own right. In so doing, we follow the Linked

Data¹ approach, which supports descriptions (of non-information resources) as first-class citizens.

We note that content and descriptions are independent from each other: an object can have content without having any description, or vice versa. We also note that each description is defined according to a schema, and that there are several standard schemas today according to which descriptions can be defined (*e.g.*, Dublin Core [8], CIDOC CRM [5], and others). For our purposes, we view a schema as consisting of a set of classes and a set of properties, organized into is-a hierarchies; moreover, a property has one or more domains and one or more ranges. We note that classes and properties are well-known concepts in object-oriented modelling, also widely used in Description Logics [2] and adopted in the semantic web framework through RDFS [10].

We finally note that when defining a description, a common practice in DLs is to use classes and properties coming from one or more standard schemas; for example, one may describe the title of a book using the title property from Dublin Core and the creation of the book using the Event class from CIDOC CRM. Our model supports this practice.

3 The Formal Definition

3.1 The Language \mathcal{L}

The language that we propose is a function-free first-order language, with the following predicate symbols:

- $\text{SchCl}(s, c)$, meaning that schema s contains class c .
- $\text{SchPr}(s, p)$, meaning that schema s contains property p .
- $\text{Dom}(s, p, c)$, meaning that in schema s property p has class c as one of its domains.
- $\text{Ran}(s, p, c)$, meaning that in schema s property p has class c as one of its ranges.
- $\text{IsaCl}(s, c_1, c_2)$, meaning that in schema s class c_1 is a sub-class of class c_2 .
- $\text{IsaPr}(s, p_1, p_2)$, meaning that in schema s property p_1 is a sub-property of property p_2 .
- $\text{SchDes}(d, s)$, meaning that d is a description over schema s . A description can only exist in association with a schema, over which it is said to be defined.
- $\text{DescCl}(d, c)$, meaning that description d describes objects that are, possibly among other things, instances of class c (hence any object described by d is an instance of class c).
- $\text{DescPr}(d, p, o)$, meaning that description d describes objects that have, possibly among other things, o as a value of property p .
- $\text{Cont}(o_1, o_2)$, meaning that object o_1 is in the content of object o_2 .
- $\text{Desc}(d, o)$, meaning that d is a description of o .

The first-order language defined on the above predicate symbols will be denoted as \mathcal{L} . Before we proceed further with the definition of the model, let's see how a real example can be represented using \mathcal{L} .

¹ <http://linkeddata.org/>

3.2 An Example

Consider the famous painting “Mona Lisa”. The painting itself is *not* a digital object, but since it is a very popular object, there exists several digital objects that identify it and that can then be seen as surrogates for it. For the present example, we use the DBpedia² identifier: http://dbpedia.org/resource/Painting_Mona_Lisa, which we abbreviate as *ml* for convenience.

As a description of Mona Lisa, we will use the one found on the Joconde database of the French Ministry of Culture³. We use as identifier of this description its URL, which we abbreviate as *d* for convenience. The description is given in Figure 1 (but please refer also to the web site for the colour codes).

In order to represent this description in our model, every field (in orange in Fig. 1) is modeled as a property in *d*; the value of each field is modelled as a value of the corresponding property *p*. Since we do not have direct access to the Joconde database containing *d*, we assume that non-clickable values (in black) are strings, even though their internal representations may be quite different. The clickable values (in brown) are digital objects in our model: they have content and such content is made accessible by the browser via clicks. Browsers that comply with the Linked Data method⁴ can be understood in terms of our model as making also descriptions of digital objects accessible via clicks (through redirections, cool URIs or other means).

For simplicity, we use the name of each field as the corresponding property; for values, we represent strings by enclosing them between quotes, and digital objects as the values themselves prefixed by the symbol “&”. Based on these conventions, the first and third fields from the top are represented as follows:

```
DescPr(d, Domaine, &peinture)
DescPr(d, Titre, “PORTRAIT DE MONA LISA...”)
```

For the sake of the example, we understand the field *Type d’object* as giving the class of which the painting is an instance. This is represented in the model as:

```
DescCl(d, tableau)
```

The description includes three thumbnails. Each thumbnail is a digital object, identified through its URI. For convenience, let t_1 , t_2 and t_3 denote the identifiers of these thumbnails. Application of the “viewing function” on each t_i would yield the image (due to lack of space, the viewing function is not discussed in the present paper). Each thumbnail can be associated to the description *d* via a special field, which we call *thumbnail*. So we have:

```
DescPr(d, thumbnail,  $t_1$ )
```

² <http://wiki.dbpedia.org/About>

³ http://www.culture.gouv.fr/public/mistral/joconde_fr?ACTION=CHERCHER&FIELD_1=REF&VALUE_1=000PE025604

⁴ <http://linkeddata.org/>

Réponse n° 1	Domaine peinture
	Type d'objet tableau
	Titre PORTRAIT DE MONA LISA (1479-1528) ; DITE LA JOCONDE
	Auteur/exécutant LEONARDO DI SER PIERO DA VINCI ; VINCI Léonard de (dit)
	Précision auteur/exécutant Vinci, 1452 ; Amboise, 1519
	Ecole Italie
	Période création/exécution 1er quart 16e siècle
	Milésime création/exécution 1503 entre ; 1506 et
	Genèse oeuvre en rapport ; reproduit en gravure
	Historique commandé par le florentin Francesco del Giocondo, époux de Mona Lisa entre 1503 et 1506 ; nombreuses copies dont une conservée au Louvre ; gravé par Fauchery, par Filhol, par Landon
	Matériaux/techniques peinture à l'huile ; bois
	Mesures 77 H ; 53 L
	Sujet représenté portrait (Mona Lisa, femme, à mi-corps, de trois-quarts, assis, accoudé, loggia, italien) ; fond de paysage (montagne, rocher, cours d'eau, pont, plaine, route)
	Date sujet représenté 1479-1528
	Lieu de conservation Paris ; musée du Louvre département des Peintures
	 Musée de France au sens de la loi n°2002-5 du 4 janvier 2002
	Statut juridique propriété de l'Etat ; musée du Louvre département des Peintures
	Anciennes appartenances François Ier ; Couronne de France
	Numéro d'inventaire INV 779
	Commentaires légère diminution du tableau sur les côtés (environ 7 mm) ; acheté vraisemblablement vers 1519, après la mort de l'artiste
	Bibliographie HEYDENRICH 6 ; OTTINO DELLA CHIESA 31 ; VILLOT I 484 ; HAUTECOEUR 1601 ; C.S.I. 1981, P 192
	Copyright notice © Musée du Louvre, © Direction des Musées de France, 1999
	Credits photographiques © Réunion des musées nationaux ; © Hervé Lewandowski ; © Thierry Le Mage
	 commande reproduction et/ou conditions d'utilisation
	renseignements sur le musée
	000PE025604

Fig. 1. A description of Mona Lisa

and the same for the other two thumbnails. Finally, we assert that d is indeed a description of Mona Lisa (*i.e.* of ml). This is done by the formula:

$$\text{Desc}(d, ml)$$

3.3 The Set of Axioms \mathcal{A}

Axioms capture the meaning of the predicate symbols, thereby constraining the interpretations of the theory to those that respect the meaning. In the following, every axiom of our theory is first stated in natural language, then it is stated formally. All variables are universally quantified.

- (A1) If property p has class c as one of its domains in a schema s , then s must contain both p and c : $\text{Dom}(s, p, c) \rightarrow (\text{SchPr}(s, p) \wedge \text{SchCl}(s, c))$
- (A2) If a property p has class c as one of its ranges in a schema s , then s must contain both p and c : $\text{Ran}(s, p, c) \rightarrow (\text{SchPr}(s, p) \wedge \text{SchCl}(s, c))$
- (A3) If c_1 is a sub-class of c_2 in a schema s then s must contain both c_1 and c_2 : $\text{IsaCl}(c_1, c_2, s) \rightarrow (\text{SchCl}(s, c_1) \wedge \text{SchCl}(s, c_2))$
- (A4) If p_1 is a sub-property of p_2 in a schema s then s must contain both p_1 and p_2 : $\text{IsaPr}(p_1, p_2, s) \rightarrow (\text{SchPr}(s, p_1) \wedge \text{SchPr}(s, p_2))$
- (A5) If a description d contains a class c , then c must be contained in the schema of d : $\text{DescCl}(d, c) \wedge \text{SchDes}(d, s) \rightarrow \text{SchCl}(s, c)$
- (A6) If a description d defines object o as a p -value, then p must be contained in the schema of d : $\text{DescPr}(d, p, o) \wedge \text{SchDes}(d, s) \rightarrow \text{SchPr}(s, p)$

(A7) If c_1 is a class in d , d is defined over the schema s and c_1 is a sub-class of c_2 in s , then also c_2 is a class in d : $(\text{DescCl}(d, c_1) \wedge \text{SchDes}(d, s) \wedge \text{IsaCl}(s, c_1, c_2)) \rightarrow \text{DescCl}(d, c_2)$

(A8) If d defines o as a p_1 -value and p_1 is a sub-property of p_2 in the schema of d , then d also defines o as a p_2 -value:

$$(\text{DescPr}(d, p_1, o) \wedge \text{SchDes}(d, s) \wedge \text{IsaPr}(s, p_1, p_2)) \rightarrow \text{DescPr}(d, p_2, o)$$

(A9) If d defines object o as a p -value, s is the schema of d and class c is one of the domains of p in s , then also c is a class in d :

$$(\text{DescPr}(d, p, o) \wedge \text{SchDes}(d, s) \wedge \text{Dom}(s, p, c)) \rightarrow \text{DescCl}(d, c)$$

We shall denote the above set of axioms as \mathcal{A} , and we shall refer to the first-order theory defined by \mathcal{L} and \mathcal{A} as the theory \mathcal{T} .

As customary, an interpretation of the language \mathcal{L} is a pair (D, I) where D is the domain of interpretation and I is the interpretation function, assigning a relation of appropriate arity over D to each predicate symbol in \mathcal{L} . To define a DL over \mathcal{L} , we consider interpretations of \mathcal{L} of a particular kind, namely interpretations having the set \mathbf{O} of digital objects as domain of interpretation (*i.e.*, $D = \mathbf{O}$). A model of \mathcal{T} is defined to be any interpretation of \mathcal{L} that satisfies all axioms in \mathcal{A} .

Intuitively, in a DL, an interpretation I is created by the facts inserted by users when they record information about objects, their contents and their descriptions. The resulting DL is then given by applying the theory to these facts. In order to make this concept precise, we re-write the axioms of our theory in the form of a positive datalog program $P_{\mathcal{A}}$ as follows:

$$\begin{aligned} \text{SchPr}(s, p) &:- \text{Dom}(s, p, c) \\ \text{SchCl}(s, c) &:- \text{Dom}(s, p, c) \\ \text{SchPr}(s, p) &:- \text{Ran}(s, p, c) \\ \text{SchCl}(s, c) &:- \text{Ran}(s, p, c) \\ \text{SchCl}(s, c_1) &:- \text{IsaCl}(c_1, c_2, s) \\ \text{SchCl}(s, c_2) &:- \text{IsaCl}(c_1, c_2, s) \\ \text{SchPr}(s, p_1) &:- \text{IsaPr}(p_1, p_2, s) \\ \text{SchPr}(s, p_2) &:- \text{IsaPr}(p_1, p_2, s) \\ \text{SchCl}(s, c) &:- \text{DescCl}(d, c), \text{SchDes}(d, s) \\ \text{SchPr}(s, p) &:- \text{DescPr}(d, p, o), \text{SchDes}(d, s) \\ \text{DescCl}(d, c_2) &:- \text{DescCl}(d, c_1), \text{SchDes}(d, s), \text{IsaCl}(s, c_1, c_2) \\ \text{DescPr}(d, p_2, o) &:- \text{DescPr}(d, p_1, o), \text{SchDes}(d, s), \text{IsaPr}(s, p_1, p_2) \\ \text{DescCl}(d, c) &:- \text{DescPr}(d, p, o), \text{SchDes}(d, s), \text{Dom}(s, p, c) \end{aligned}$$

Given an interpretation I , the above rules will be applied to I in order to derive the minimal model of $P_{\mathcal{A}}$ containing I . This application is expressed by the immediate consequence operator $T_{P_{\mathcal{A}}}$, which is well-known to be monotone and therefore it admits a minimal fix-point $\mathcal{M}(P_{\mathcal{A}}, I)$. For more details, see [9].

Definition 1 (Digital Library). Let I be any interpretation of \mathcal{L} . We call digital library over I , denoted DL_I , the minimal model $\mathcal{M}(P_{\mathcal{A}}, I)$ of \mathcal{A} that contains I .

We note that in our definition of DL we have used the *same* domain for the interpretation of *all* predicates, namely the set of digital objects \mathcal{O} . The use of a common domain implies that users inserting facts in the DL may (knowingly or unknowingly) end up using the same object with different roles; for example, one can state that a schema named *Rome* has a property also named *Rome*, by instantiating the predicate $SchPr$ as follows:

$$SchPr(Rome, Rome)$$

Such a statement can be treated by the theory without any problem— although using different names for the schema and the property would be a better way of modeling information. Preventing the use of same name for different roles is a good modeling practice; however, we believe that this issue should not be treated at the theory level (*i.e.*, by adding extra axioms), but rather at the implementation level by putting in place mechanisms to guide users in this respect.

4 Querying a Digital Library

When searching a DL, descriptions may become obstacles in between users and objects, in a very real sense. For instance, when searching for objects about *lattices* authored by *John*, the user would have to use a query like:

$$(\exists d_1 d_2) Desc(d_1, x) \wedge Desc(d_2, x) \wedge DescPr(d_1, author, John) \wedge DescPr(d_2, about, lattices)$$

This query is unnecessarily cumbersome, as it mentions two descriptions d_1 and d_2 which have nothing to do with the information needed by the user. A more intuitive and straightforward way of expressing the user’s information need is to relate authorship and aboutness directly to the sought objects.

In order to simplify the expression of queries, we introduce two additional predicate symbols that allow to directly connect description elements with the objects they are associated with. These predicates are:

- $CIExt(c, o)$, meaning that object o is an instance of class c .
- $PrExt(o_1, p, o_2)$, meaning that object o_1 has object o_2 as p -value.

Using the latter predicate symbol, the previous query can be expressed as follows:

$$PrExt(x, author, John) \wedge PrExt(x, about, lattices)$$

Clearly, this is a direct translation of the user’s information need. In view of this example, we augment the language \mathcal{L} by adding the predicate symbols $CIExt$ and $PrExt$ to it. We call \mathcal{L}_+ the resulting first-order language.

Now, in order to define the semantics of these two additional predicates, we consider that object o is an instance of class c if there exists some description d establishing this relationship. This may happen in one of two different ways:

- c is a class in d and d is a description of o .
- c is one of the ranges of a property p in the schema of d , and d defines o as a p -value.

Analogously, object o_2 is a p -value of object o_1 if o_1 has a description d that defines o_2 to be a p -value. More formally, the predicate symbols CIEExt and PrExt are related to the predicate symbols in \mathcal{L} by the following axioms:

(Q1) $\text{Desc}(d, o) \wedge \text{DescCl}(d, c) \rightarrow \text{CIEExt}(c, o)$

(Q2) $\text{Desc}(d, o_1) \wedge \text{SchDes}(d, s) \wedge \text{Ran}(s, p, c) \wedge \text{DescPr}(d, p, o_2) \rightarrow \text{CIEExt}(c, o_2)$

(Q3) $\text{Desc}(d, o_1) \wedge \text{DescPr}(d, p, o_2) \rightarrow \text{PrExt}(o_1, p, o_2)$

We shall denote as \mathcal{A}_+ the set of axioms in \mathcal{A} augmented by the above axioms. Axioms Q1, Q2 and Q3 can be translated into the following, equivalent positive datalog rules:

$\text{CIEExt}(c, o) :- \text{Desc}(d, o), \text{DescCl}(d, c)$

$\text{CIEExt}(c, o_2) :- \text{Desc}(d, o_1), \text{SchDes}(d, s), \text{range}(d, p, o_2), \text{DescPr}(d, p, o_2)$

$\text{PrExt}(o_1, p, o_2) :- \text{Desc}(d, o_1), \text{DescPr}(d, p, o_2)$

If we add these rules to the datalog program $P_{\mathcal{A}}$ seen earlier, then we obtain a positive datalog program $P_{\mathcal{A}_+}$, which is equivalent to the axioms in \mathcal{A}_+ .

Definition 2 (Query over a digital library). A query over a digital library is any open well-formed formula $\alpha(x_1, \dots, x_n)$ of \mathcal{L}_+ with $n \geq 1$ free variables x_1, \dots, x_n .

Intuitively, the answer of a query with n free variables is the set of n -tuples of objects $\langle o_1, \dots, o_n \rangle$ such that, when every variable x_i is bound to the corresponding object o_i , the resulting ground formula of \mathcal{L} is true in DL_I . Formally, we have the following definition.

Definition 3 (Answer of a query). The answer of a query $\alpha(x_1, \dots, x_n)$ over a digital library DL_I is given by:

$$\text{ans}(\alpha, I) = \{ \langle o_1, \dots, o_n \rangle \mid \alpha(o_1, \dots, o_n) \in \mathcal{M}(P_{\mathcal{A}_+}, I) \}$$

To exemplify, let us consider again our earlier example, where Mona Lisa is represented by the object ml and described by the description in Figure 1 (which is identified as d , and therefore $\text{Desc}(d, ml)$ is in the DL). As a consequence of this assertion (and of axioms Q1, Q2 and Q3), we have:

$\text{CIEExt}(ml, \text{tableau})$

$\text{PrExt}(ml, \text{Domaine}, \&\text{peinture})$

$\text{PrExt}(ml, \text{Titre}, \text{"PORTRAIT DE MONA LISA..."})$

These last statements allow a user to discover the Mona Lisa painting as a *tableau* via the query:

$\text{CIEExt}(x, \text{tableau})$

Alternatively, Mona Lisa can be discovered as an object whose domain is *peinture* via the query:

$\text{PrExt}(x, \text{Domaine}, \&\text{peinture})$

5 Implementation Issues

Regarding the implementation of our model, a straightforward way is to use a relational database. This is a most natural choice for two reasons: first, an interpretation of \mathcal{L}_+ consists just of a set of relations that can be implemented as tables; second, any predicate in a logic defined on \mathcal{L}_+ can be translated in a straightforward manner to an SQL WHERE clause on an interpretation of \mathcal{L}_+ . More importantly, by choosing an implementation strategy based on relational technology, we can take advantage of the scalability and the optimized query evaluation of relational DBMSs, at a minimal effort.

A simple strategy for evaluating queries against a DL could then consist of the following steps:

1. Store the initial set of facts I in a relational database $RDB(I)$; as already noted, the correspondence between I and $RDB(I)$ is straightforward.
2. Expand $RDB(I)$ until obtaining the database $RDB(\mathcal{M}(P_{A+}, I))$ that contains the model for answering queries; this requires adding tuples to the tables in $RDB(I)$ using the inference mechanism described earlier.
3. Map each query q against the DL to an equivalent SQL query $SQL(q)$; as already noted, this mapping is straightforward.
4. Evaluate $SQL(q)$ against $RDB(\mathcal{M}(P_{A+}, I))$.

One of the problems here is to design optimal algorithms for maintaining $RDB(DL_I)$ in presence of user updates. An alternative strategy for evaluating queries against a DL would be to compute query answers directly on $RDB(I)$ *without* expanding it to $RDB(\mathcal{M}(P_{A+}, I))$. In this case, the main issue is to define an inference mechanism for answering queries directly from $RDB(I)$.

5.1 RDF and Our Model

Another implementation option would be to rely on the Resource Description Framework (RDF) [10]. RDF is a knowledge representation language for describing web resources based on notions borrowed from the web architecture, such as URIs. RDF is widely used in the context of DLs as it provides an easy to use notation for describing objects via properties. Moreover, RDF is strongly tied to the Web information space, thus providing a useful means for information sharing and integration. The basic issue in considering RDF for implementing our model, is that RDF can only represent *binary* information. On the other hand, in our language we have one predicate symbol of arity 3. Of course, instances of this symbol can be represented in RDF [1]; however, their representation requires a transformation that shifts the emphasis from the original objects to objects of a different kind, making the extraction of information cumbersome. To see this, let us consider the formula $\text{DescPr}(d, p, o)$ in \mathcal{L}_+ , asserting that description d defines object o as a p -value. In order to know the pairs of descriptions assigning the same value to the same property, the following query can be used:

$$(\exists p)(\exists o)\text{DescPr}(d, p, o) \wedge \text{DescPr}(d', p, o)$$

where d and d' represent the sought descriptions. By introducing an additional URI, i , we can represent the assertion $\text{DescPr}(d, p, o)$ via three RDF triples:

$$(i \text{ Descr } d), \quad (i \text{ Prop } p), \quad (i \text{ Obj } o)$$

Here, Descr , Prop and Obj are properties linking i respectively to the description d , the property p and the object o in the above triples. Now, each of the above triples does not describe a “real” resource (such as object o), but an artificial resource i , created for the purposes of encoding the original information in triples. Reasoning as above, we can represent $\text{DescPr}(d', p, o)$ as:

$$(j \text{ Descr } d'), \quad (j \text{ Prop } p), \quad (j \text{ Obj } o)$$

In order to extract from the triple-based representation which pairs of descriptions d and d' share the same property p and value o , we have to use a much more complex query, which reconstructs the original predicates and then relates them properly. Using an intuitive triple-based notation, such a query would be:

$$(\exists i p o j)(i \text{ Descr } d) \wedge (i \text{ Prop } p) \wedge (i \text{ Obj } o) \wedge (j \text{ Descr } d') \wedge (j \text{ Prop } p) \wedge (j \text{ Obj } o)$$

This query is much more involved than the previous one, making the extraction of information from an RDF representation cumbersome.

In summary, descriptions (as understood in the present model), are not naturally modelled in RDF. We are currently working on an RDF vocabulary that allows to capture descriptions both at the syntactical and at the semantical level, *i.e.* licensing the inferences outlined in the previous sections.

We note that RDF offers reification as a mechanism for representing a triple as an object. Via reification, a more general information model is obtained, able to cope with ternary predicates. Unfortunately, reification does not have any semantics attached to it [7], thus the information encoded in reified triples cannot be exploited, *e.g.*, for discovery purposes. Moreover, all problems remain for predicates with arity higher than 2.

6 Other Related Work

The model presented in this paper has drawn major inspiration from the DELOS reference model for DLs (DRM for short) [43]. The DRM has been the result of a large effort, mainly conducted within the DELOS Network of Excellence [5]; it aims at providing an ontology for DLs, defining all relevant concepts in an informal, yet precise language. These concepts are grouped in the DRM under six main categories: content, user, functionality, quality, policy and architecture. The DRM content category includes descriptions and their formats. The present work can be seen as an attempt to formalize the DRM content dimension, providing at the same time a basis for an implementation.

Another remarkable DL model is the 5S model [6]. Like DRM, the 5S model aims at capturing all aspects of a DL, from content (Streams making up Structures, in turn defining Spaces) to users (Societies) and usages (Scenarios). The

⁵ <http://www.delos.info>.

basic difference between the 5S model and our model is that 5S aims at generality and coverage, while being at a too abstract level to immediately suggest an implementation. Our model, on the other hand, is more focused, leaving out users and usages and addressing only the content dimension. As a result we have a smaller and simpler set of definitions, which nonetheless include a query language and can be used to design a DL management system at the level of complexity required, for example, by Europeana.

7 Conclusions

We have introduced a first-order language for describing DLs consisting of a set of digital objects, with two fundamental features: content and descriptions. Each of these features is expressed in the language via a certain set of predicates, the semantics of which is fixed by the axioms of the theory. A DL is then defined as one particular type of model of the underlying theory, the existence and uniqueness of which has been proved by reducing the theory to an equivalent datalog program. Two additional predicate symbols have been introduced to ease query expression, and the underlying theory has been extended to deal with the semantics of these symbols. Implementation has been briefly discussed, highlighting the necessity of extending RDF in order to capture the semantics of the model.

Acknowledgements. The work reported in this paper has been partially supported by the PICS program and the ASSETS project.

References

1. Defining n-ary relations on the semantic web, W3C Working Group Note (April 2006), <http://www.w3.org/TR/swbp-n-aryRelations/>
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): *The Description Logic Handbook: Theory, Implementation, and Applications*, 2nd edn. Cambridge University Press, Cambridge (2003)
3. Candela, L., Castelli, D., Ioannidis, Y., Koutrika, G., Pagano, P., Ross, S., Schek, H.-J., Schuldt, H.: *Setting the foundations of digital libraries the delos manifesto*. *D-Lib Magazine* 13(3/4) (March/April 2007)
4. Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Ioannidis, Y., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobрева, M., Katifori, V., Schuldt, H.: *The DELOS Digital Library Reference Model - Foundations for Digital Libraries*. In: *DELOS Network of Excellence on Digital Libraries (2007)*, ISBN 2-912337-37-X
5. Doerr, M.: *The CIDOC conceptual reference model: An ontological approach to semantic interoperability of metadata*. *AI Magazine* 24(3), 75–92 (2003)
6. Goncalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: *Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries*. *ACM Transactions on Information Systems* 22(2), 270–312 (2004)
7. Hayes, P.: *RDF Semantics*. W3C Recommendation, WWW Consortium (February 2004), <http://www.w3.org/TR/rdf-mt/>

8. Dublin Core Metadata Initiative. Dublin core metadata initiative dublin core metadata element set, version 1.1 (January 2008), <http://dublincore.org/documents/dces/>
9. Lloyd, J.W.: Foundations of Logic Programming. Springer, Heidelberg (1987)
10. Manola, F., Miller, E.: RDF Primer. In: W3C Recommendation, WWW Consortium (February 2004), <http://www.w3.org/TR/rdf-primer/>
11. Meghini, C., Spyratos, N.: Modelling the web. Invited talk at the Fourth Franco-Japanese Workshop ISIP: Information Search, Integration and Personalization, Paris, France (October 6-8, 2008)
12. Meghini, C., Spyratos, N.: Rationale and some principles for a VLDL data model. Invited talk at the Very Large Digital Libraries (VLDL 2008) Workshop, held in conjunction with ECDL 2008, Aarhus, Denmark (September 2008)
13. Meghini, C., Spyratos, N., Yang, J.: A data model for digital libraries. Invited talk at the Max Planck eScience Seminar on Repository Systems, Garching, Germany (June 2009)

General-Purpose Digital Library Content Laboratory Systems

Paolo Manghi, Marko Mikulicic, Leonardo Candela,
Michele Artini, and Alessia Bardi

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1 - 56124 Pisa - Italy
name.surname@isti.cnr.it

Abstract. In this work, we name Digital Library Content Laboratories (DLCLs) software systems specially devised for aggregating and elaborating over information objects – e.g., publications, experimental data, multimedia and compound objects – collected from possibly heterogeneous and autonomous data sources. We present a general-purpose and cost-efficient system for the construction of customized DLCLs, based on the D-NET Software Toolkit. D-NET offers a service-oriented framework, where developers can choose the set of services they need, customize them to match domain requirements, and combine them in a “LEGO fashion” to obtain a personalized DLCL. D-NET is currently the enabling software of several DLCLs, operated by European Commission projects and national initiatives.

1 Introduction

In the last decade digital library systems evolved moving from stand-alone systems serving one community of users, to possibly distributed platforms offering rich content across several research communities. To meet such requirements, a number of organizations, from research and academic institutions to national consortia, invested in digital library management systems specially devised for manipulating *information objects* [2] – publications, audio and video material, experimental data sets, “compound objects”, etc., and “surrogates” of all these, i.e., metadata representations of information objects – collected from a set of possibly heterogeneous and autonomous data sources. Renown examples of such systems can be found in the institutional repository area, where research communities are interested in processing publications (e.g., OCLC-OAster,¹ BASE,² DAREnet-NARCIS³), and lately experimental data, collected from OAI-PMH data sources; or in projects such as SAPIR⁴, where an advanced system was built to automatically extract indexing features from images and videos collected from web sources. Although distinct in the nature of the information objects they handle, such systems have common functional and architectural patterns

¹ OCLC OAster, <http://www.oaister.org>

² BASE: Bielefeld Academic Search Engine, <http://www.base-search.net>

³ DAREnet: Digital Academic Repositories, <http://www.narcis.nl>

⁴ Search In Audio Visual Content Using Peer-to-peer IR, <http://www.sapir.eu>

regarding the *collection, storage, manipulation, and provision* of information objects. They share architectural issues such as scalability (e.g., support access to large sets of objects), robustness (e.g., preserve objects from destruction), and administration of a set of data sources, hence deal with tasks such as data workflow definition and scheduling (e.g., weekly collection, storage and manipulation of objects from data sources). In the following we refer to these as *Digital Library Content Laboratories* (DLCLs), to capture the essence of their core functionalities.

In this paper, we present a novel general-purpose DLCL system, which developers can exploit to construct customized DLCLs in a cost-effective way. The solution is based on the *D-NET Software Toolkit* [1], a service-oriented application framework developed in the context of the DRIVER and DRIVER-II projects.⁵ D-NET delivers advanced data management services which can be configured and combined in a LEGO-like approach to form personalized data workflows and construct customized DLCLs. D-NET's key properties are: (i) the *customizability, extensibility, modularity* of its functionalities, to support customization of DLCL functionality, and (ii) *autonomicity, distribution and sharing* of its instantiations, to enable the operation of "intelligent", scalable and robust DLCL systems.

The paper is organized as follows: Section 2 presents a high-level architecture for general-purpose DLCLs, resulted from rationalizing the issues that practitioners in the field have to face when designing and developing such systems; Section 3 presents the general-concepts of D-NET; Section 4 describes D-NET's data management services and how customized, robust and scalable DLCLs can be constructed out of them; finally, Section 5 concludes the paper.

2 General-Purpose Digital Library Content Laboratory Systems

As mentioned in the introduction, by Digital Library Content Laboratory (DLCL) we mean a software system whose components address the core functionalities of collecting and store information objects from a set of data sources (e.g., repositories, archives, libraries, databases) so as to manipulate and provide them to third party applications. Depending on the application domain, DLCL components deal with information objects which may range from *metadata records* (e.g., Dublin Core⁶ records, MARCXML⁷ records, ESE records⁸) and *payloads* (e.g., video files, text files, image files) to *compound objects*, here intended as graphs of information objects connected by named relationships. In the following, to abstract from the specific domain, we assume that DLCL components deal with information objects of a given *data model* and that data models can be: *metadata data models*, i.e., metadata formats, *payload data models*, i.e., file formats, or *compound object data models*, i.e., structure graphs representing the structure and semantics of the given compound objects. Based on this high-level vision, a DLCL can be conceived as a set of components organized in four functional areas (**Fig. 1**):

⁵ *Digital Repository Infrastructure Vision for European Research*, <http://www.driver-community.eu>

⁶ *Dublin Core Metadata Initiative*, <http://dublincore.org>

⁷ *MARC XML: the MARC 21 XML Schema*, <http://www.loc.gov/standards/marcxml>

⁸ *ESE XML Schema*: <http://www.europeana.eu/schemas/ese/>

Data storage: components in this area manage storage and access for a collection of information objects conforming to a given data model. Examples are: an index component handling Dublin Core metadata objects (data model: Dublin Core metadata objects) or a “conference proceedings management” component, managing compound objects representing books of proceedings connected with relationships to the respective article payload objects and metadata objects (data model: “proceedings” compound objects).

Data mediation: components in this area are capable of managing a set of available “external” data sources and of serving requests for exchanging (retrieve and/or deliver) information objects of a given data model with them. The exchange is based on standard APIs (e.g., OAI-PMH, OAI-ORE, FTP, WSDL/SOAP, SRW, REST). For example, a component for administrating a federation of library catalogues, exposing JDBC APIs, whose participants may join or leave the federation any time.

Data provision: components in this area allow third-party applications to access information objects collections in the storage area according to standard APIs (e.g., OAI-PMH, OAI-ORE, FTP, WSDL/SOAP, SRW, REST). Examples are: OAI-PMH harvesters willing to access DLCL metadata objects collections; a web portal, exposing the information objects in the DLCL which are accessible through SRW APIs.

Data manipulation: components in this area offer functionality for processing information objects. Examples are: transformation of information objects from one data model into another (e.g., from DOC payload objects to PDF’s, from MARC metadata objects to Dublin Core’s), extraction of information from information objects (e.g., extracting histograms from image payloads), validation quality of information objects.

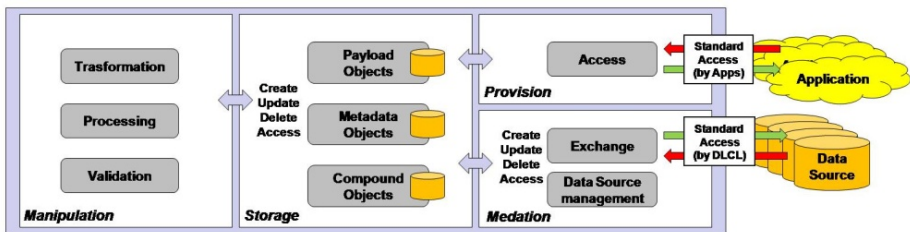


Fig. 1. DLCLs: functional architecture

General-purpose DLCL systems are highly adaptable DLCLs capable of meeting DLCL requirements of arbitrary user communities. To this aim, such systems and their components should be designed according to the following principles:

Modularity: components should provide minimal units of functionality so that they can be arbitrarily composed to meet custom data management workflows. With reference to the example above, a Repository Aggregator System may harvest metadata objects, store them and then index them, while another one may instead need to harvest and then index them straightaway.

Customizability: components should support polymorphic functionalities, operating over information objects whose data model matches a generic structural template.

For example, a metadata object indexing component should be designed to be customizable to any metadata object data model.

Extensibility: the system should be open to the addition of new components, in order to introduce new functionality, whenever this is required and without compromising the usability of other components.

3 D-NET in a Nutshell

The *D-NET Software Toolkit* was designed and developed within the DRIVER and DRIVER-II EC projects. Its software is open source, developed in Java and available for download [1]. D-NET implements a Service Oriented Architecture (SOA), based on the Web Service framework,⁹ capable of operating run-time environments where *multiple organizations* can share data sources and services to *collaboratively* construct applications for creating new information object collections, and deliver them through standard APIs to third-party applications. To this aim, D-NET provides several services from which organization developers can choose to assemble their DLCL applications. Specifically, a D-NET infrastructure is made of two main logical layers: the system core, called *enabling layer*, whose function is to support the operation of the *application layer*, which consists of the services forming the DLCLs. The enabling layer comprises four services:

Information Service (IS). The service addresses *registration* and *discovery* of services. As in peer-to-peer systems, developers can dynamically add or remove services from the system, that is from the DLCLs they operate. To this aim, services *register* to the IS their *profile* (information about their location, the functionality they expose and their current status) and *discover* the services they need by searching profiles in the IS. The IS software is built on *eXist-db*, an Open Source native XML database.¹⁰

Manager Service (MS). The service addresses service *orchestration*. The MS can be configured by developers to autonomously execute *workflows*, i.e., distributed transactions, involving a number of services. Typically, the MS reacts to an event, locates through the IS the services needed for the relative workflow, instructs them and monitors their behavior. The MS implements a graph-based workflow engine based on the *Sarasvati* project.¹¹

Authentication and Authorization Service (AAS). The service addresses AA communication, i.e., services may concede access only to Authorized services and users; AA policies are expressed through the XACML standard.

ResultSet Service. The service manages *ResultSets*, i.e., “containers” for transferring list of objects between a “provider” service and a “consumer” service. Technically, a ResultSet is an ordered lists of files identified by an *End Point Reference (EPR)*,¹² which can be accessed by a consumer through paging mechanisms, while being fed by a provider. D-NET services can be designed to accept or return ResultSet EPRs as input parameters or results to invocations, in order to reduce response delays and limit the objects to be transferred.

⁹ W3C Web Services Activity web site, <http://www.w3.org/2002/ws>

¹⁰ *eXist-db*, <http://exist.sourceforge.net>

¹¹ *Sarasvati project*, <http://code.google.com/p/sarasvati>

¹² The Web Service EPR standard describes the location of a resource on the internet.

4 Constructing Customized DLCLs in D-NET

The D-NET framework provides a *Data Management service kit*, whose services implement typical DLCL components. In this section, we highlight the properties of customizability, extensibility and modularity of these services, which reflect the architectural desiderata of general-purpose DLCLs. Secondly, we present how developers can assemble scalable, robust and intelligent DLCLs in a D-NET system, exploiting service autonomicity, distribution and sharing.

4.1 Data Management Kit

In this section we introduce the services available in the latest D-NET release (v1.2) of the Data Management kit [1], organized in the four DLCL's architecture functional areas (see Fig. 2).

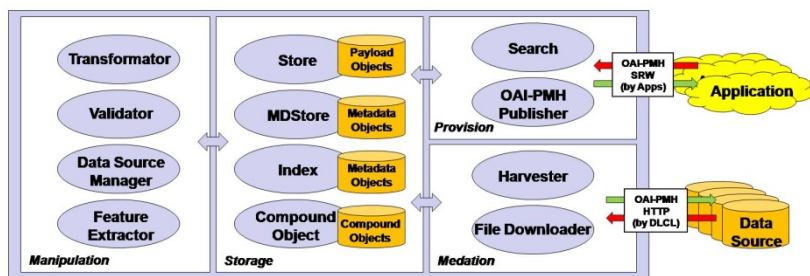


Fig. 2. DLCLs in D-NET: functional architecture

Such services are designed following principles of customizability and modularity and exchange long lists of information objects through the *ResultSet* mechanism. In particular, objects to be exchanged are serialized through XML representations, called *Object Representations* (OR). The XML of an OR has an *header* section with the unique identifier and the name (unique in the system) of the data model of the object, and a *body* section, which incorporates the XML representation of the object. In the following, when mentioning information objects transferred over-the-wire through *ResultSets*, we implicitly refer to their OR serialization.

Data Storage. Services in this area manage payload, metadata and compound objects of any data models. Bulk transfer of objects to a given service is performed by sending a *ResultSet* EPR with the objects to service, which will actively pull the objects from the *ResultSet* for local usage.

MDStore Service. An MDStore Service manages a set of *MDStore units* capable of storing metadata objects of a given metadata data model. Consumers can create and delete units, and add, remove, update, fetch, get statistics on metadata objects from-to a given unit.

Index Service. An Index Service manages a set of *Index units* capable of indexing metadata objects of a given data model and replying CQL queries¹³ over such objects. Consumers can feed units with records, remove records or query the records. The Service is implemented on *Yadda*¹⁴, a software project based on *Apache Lucene* index.¹⁵

¹³ *Contextual Query Language*, <http://www.loc.gov/standards/sru/specs/cql.html>

¹⁴ *Yadda*, <http://yaddainfo.icm.edu.pl>

¹⁵ *The Apache Lucene Project*, <http://lucene.apache.org/java/docs>

Store Service. The Store Service manages a set of *Store units* capable of storing payload objects of a given data model. Consumers can store and fetch payload objects from a Store unit, in particular can retrieve them by identifier or by bulk retrieval (e.g., all objects, by date of creation) to be returned as a ResultSet of ORs.

Compound Object Service. The Compound Object Service offers tools to grow a graph of payload and metadata objects from Store and MDStore services by providing named relationships between the relative identifiers. Consumers can create and remove relationships and run xpath navigational queries across the graph. Query results are returned through a ResultSet that contains the OR representations of the objects matching the query. The Service builds on *DORoTy* [3], which in turn is based on the *Neo4j* project to handle triple stores.¹⁶

Data Mediation. Services in the mediation area are capable of fetching data from external data sources by means of OAI-PMH interfaces or by downloading files available at given URLs.¹⁷ The aim of such services is to retrieve external objects and convert them into corresponding OR representations of a given data model.

Harvester Service. An Harvester Service can execute the six OAI-PMH protocol verbs over a given repository data source registered to the system. The verb `ListRecords` fetches repository metadata records and returns the EPR of a ResultSet that contains their OR serialization.

File Downloader. A File Downloader Service can download (http) and deposit into a given Store unit a set of files whose URLs are provided through a ResultSet of payload objects OR.

Data Provision. Services in the data provision area interface external applications with objects in the storage area. Currently, the services offer SRW¹⁸ interfaces, to operate over the available Index Services units, and OAI-PMH interfaces, to access the metadata objects in the MDStore Service units.¹⁹

Search Service. A Search Service offers an SRW interface accepting a query Q and a metadata data model, to route and run Q over the Index units matching the model (note that such units are discovered through the Information Service). Responses, when more than one Index Service is involved, are then “fused” and pushed into a ResultSet, whose EPR is returned as result.

OAI-PMH Publisher Service. An OAI-PMH Publisher Service offers OAI-PMH interfaces to third-party applications (i.e., harvesters) willing to access metadata objects in the MDStore units.

Data Manipulation. Services in the manipulation area are capable of (i) manipulating objects to transform them from one data model to another or verify they respect certain structure and semantics, and (ii) handling a set of data sources, in order to organize and configure the respective data manipulation workflows.²⁰

Feature Extractor Service. A Feature Extractor Service generates a ResultSet of OR objects from a ResultSet of input OR objects by applying a given extraction

¹⁶ *Neo4j The Graph Database*, <http://neo4j.org>

¹⁷ Services for accessing data sources responding to OAI-ORE and ODBC interfaces are being developed for D-NET v2.0

¹⁸ *Search/Retrieval via URL*, <http://www.loc.gov/standards/sru>

¹⁹ Services for providing access to compound object service through OAI-ORE and ODBC interfaces are being developed for D-NET v2.0

²⁰ Services for providing Authority File Management and Citation Inference are being developed for D-NET v2.0.

algorithm. Examples are: extracting the histograms of image payload objects; extracting full-text of PDF payload objects; generating payload objects from payload objects (DOC to PDF). Algorithms can be plugged-in as special software modules, whose invocation becomes available to consumers.

Transformer Service. A Transformer Service is capable of transforming metadata objects of one data model into objects of one output data model. The logic of the transformation, called *mapping*, is expressed in terms of a rule language offering operations such as: (i) field removal, addition, concatenation and switch, (ii) regular expressions, (iii) invocation of an algorithm through a Feature Extractor Service, or (iv) upload of full XSLT transformations. User interfaces support administrators at defining, updating and testing a set of mappings. The result of a transformation is the EPR of a ResultSet that contains the generated metadata objects.

Validator Service. A Validator Service is used by system administrators to verify the quality of a ResultSet of OR objects against a set of validation rules, capable of capturing syntactic and semantics constraints (e.g., expected vocabulary terms for fields of metadata objects). A validation operation returns a link (or sends an email) to a feedback report, which summarizes the “level of conformity” of the collection.

Data Source Manager. The service offers tools for the management of the data sources available to the system: user interfaces for the registration of external data sources and the definition of the relative orchestration workflows. To this aim, the service interacts with the DLCL Services and the Manager Services.

4.2 Assembling DLCLs in D-NET Systems

D-NET developers build customized DLCLs by selecting the services they need from the Data Management kit, customizing them to match data model and workflow requirements and registering them to the enabling layer. Besides, they might exceptionally design and develop their services to complement missing functionality. On the operational side, the service-oriented implementation of D-NET yields further properties, which provide added value to a DLCL running system:

Sharing: multiple organizations can operate their DLCLs in the same D-NET system and collaborate by sharing data sources and services. For example, one organization may provide storage and mediation services to aggregate content from a set of data sources. Such services may be shared with another organization, whose DLCL provides services to access and derive statistics from such content.

Distribution: service replicas, that is clones of functionality and content, can be kept at different sites. This makes the system more robust to network failures and system crashes (availability of service) as well as to concurrent accesses (scalability by workload distribution).

Autonomy: Manager Services can autonomously monitor and orchestrate service workflows. For example, Manager Services can be configured to guarantee that a given number of content replicas is maintained across distributed storage services in a DLCLs.

5 Conclusions

This work presented the D-NET system software as a general-purpose service-oriented framework where multiple organizations can collaboratively construct customized,

robust, scalable, intelligent and cost-effective DLCLs. D-NET is today adopted by several EC projects, national consortia and communities to create customized DLCLs under diverse application domains, and other organizations are enquiring for or are experimenting its adoption. To be mentioned are: the DRIVER project²¹ (infrastructure of Open Access European repositories), the Spanish-Recolecta repository infrastructure²² (Concorcio Madrono), the Slovenian repository infrastructure²³ (National and University Library of Ljubljana), the Belgium repository infrastructure²⁴ (University of Ghent), the EFG EC project²⁵ (European Film Archives), the OpenAIRE EC project system²⁶ (aggregation of publications funded by FP7 EC projects) and the Heritage of People in Europe EC Project system (audio/video from archives, libraries and museums of social and labor history). The D-NET software is open source and available for usage, improvement and extension by any community of developers willing to contribute.

Acknowledgments. This work would have not been possible without the precious support and cooperation of the other members of the D-NET technical team: ICM Research Centre (Warsaw) for Yadda Index Services and AA mechanisms in the Enabling Area, University of Bielefeld Library (Bielefeld, Germany) for the Data Management Area, and the Department of Informatics, National and Kapodistrian University of Athens (Greece) for the Validator Services. Research partially supported by the INFRA-2007-1.2.1 Research Infrastructures Program of the European Commission as part of the DRIVER-II project (Grant Agreement no. 212147).

References

1. D-NET Project, Istituto di Scienza e Tecnologie dell'Informazione, Centro Nazionale delle Ricerche, Pisa, Italy, ICM Research Centre, Warsaw, University of Bielefeld Library, Bielefeld, Germany and Department of Informatics, National and Kapodistrian University of Athens (Greece), <http://www.d-net.research-infrastructures.eu>
2. Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Kou-Trika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobрева, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model - Foundations for Digital Libraries. In: DELOS: a Network of Excellence on Digital Libraries (February 2008), ISSN 1818-8044, ISBN 2-912335-37-X
3. Candela, L., Manghi, P., Pagano, P.: An Architecture for Type-based Repository Systems. In: Proceedings of the Second Workshop on Foundations of Digital Libraries, in Conjunction with ECDL 2007, Budapest, Hungary, DELOS (2007)

²¹ *DRIVER infrastructure portal*, <http://search.driver.research-infrastructures.eu>

²² *Spanish infrastructure portal*, <http://search.recolecta.driver.research-infrastructures.eu>

²³ *Slovenian infrastructure portal*, <http://search.slovenia.driver.research-infrastructures.eu>

²⁴ *Belgium infrastructure portal*, <http://search.belgium.driver.research-infrastructures.eu>

²⁵ *The European Film Gateway*, <http://www.europeanfilmgateway.eu>

²⁶ *Open Access Infrastructure for Research in Europe*, <http://www.openaire.eu>

Component-Based Authoring of Complex, Petri Net-Based Digital Library Infrastructure

Yung Ah Park, Unmil Karadkar, and Richard Furuta

Center for the Study of Digital Libraries and Department of Computer Science,
Texas A&M University, College Station, TX 77843-3112, USA
caT@csdl.tamu.edu

Abstract. caT, a Petri net-based hypertext system, serves as a platform for unified modeling of digital library infrastructure and its governing policies, user characteristics, and their contextual information. Traditionally, users have created caT networks from scratch, thus limiting their use to small collections. In this paper we introduce TcAT, a component-based authoring tool, which enables the creation of large caT nets that can represent interaction-rich, real-life spaces such as libraries and museums. TcAT implements composition operations from Petri net theory, allowing authors to select and modify existing net fragments as templated building blocks for larger networks. Authors may switch between visual and textual modes at will, thus combining the strengths of expressing large nets textually and selecting net fragments via point-and-click interaction. A user evaluation of the new authoring mechanisms suggests that this is a promising tool for improving the efficiency of experienced users as well as that of novice users, who are unfamiliar with the Petri net formalism.

Keywords: caT, Petri net-based hypertext, digital library infrastructure.

1 Introduction

In addition to serving as treasure troves of information artifacts, such as books, videos, and audio materials, libraries provide spaces for social interaction as well as specialized services for their patrons' diverse needs. Trained staff helps them access library services. Sometimes, books that they desire are unavailable, as other users have borrowed these. Patrons may run into others who share their interests. The environment of libraries is dynamic and vibrant with users who share the physical space. Dominant digital library models trade the sense of space and some services for ubiquitous access to their materials. Modeling real world interactions in the digital world requires programming external to the digital library infrastructure, which can be quite expensive. Partially in response to the costs, typical digital libraries provide basic services and neglecting atypical needs or special situations.

context aware Trellis (caT) [9], a Petri-net-based hypertext system, is particularly well-suited to modeling the dynamism of traditional libraries. It enriches the ubiquity of digital access by incorporating the sense of space, referral services, and policies that govern the use of digital library materials. The caT infrastructure responds to the

characteristics of its patrons’ physical and contextual environments. It includes mechanisms for providing time-sensitive help, such as live chat with a librarian during office hours and access to an automated help system after hours [7]. caT libraries can tailor their content to the characteristics of patrons’ information devices, for example, presenting richer videos to those connected via desktop computers with high speed internet than those viewing them on mobile devices. caT supports these features by virtue of inherent properties, without requiring specialized programming.

Traditionally, caT authors have designed hypertexts (called nets) that instantiate the structure and behavior of their library element-by-element, using a visual interface that supports point-and-click interaction. While the interaction mode is familiar to the authors the process is slow and laborious when creating large networks.

In this paper, we describe advanced compositional mechanisms that aid authors in thinking in terms of patterns and processes than assembling net elements. We introduce techniques that enable authors to select and modify sample specifications to suit new requirements. Manual authoring of large structures can result in inadvertent creation of redundant net components. Our suite includes tools for simplifying net structures as another avenue for managing complex networks. Finally, we introduce textual authoring as an alternate modality to enable easy manipulation of net components. A preliminary evaluation indicates that the new features help authors create caT networks more efficiently.

The rest of this paper is organized as follows: Section 2 introduces caT. Section 3 focuses on the template-based composition mechanism for easier authoring of complex hypertexts. Section 4 discusses the results of a user evaluation of our new authoring tool. Section 5 concludes the paper with pointers to future work.

2 Context-Aware Trellis

context-aware Trellis (caT) uses colored Petri nets to specify hypertext documents. When modeling a digital library using caT, information content of the library infrastructure is associated with “places” (circles) and the actions available to patrons are called “transitions” (rectangles), as shown in fig. 1. Unlike Web pages, which embed links within the information content, caT separates the information content from the actions. Directed arcs (arrows) indicate the browsing direction and colored tokens (dots) represent users’ current locations in the hypertext structure. Users view the information from all places where their tokens exist and may browse the library by

“firing” any transitions that follow these places to continue browsing to another place. Transition activation and firing conditions, called browsing semantics, are determined by rules written in the transition and its connecting arcs [10]. Authors allow access to certain places or restrict it under specific conditions by stating appropriate browsing semantics. Information may be displayed or hidden depending upon time, day of the week, user location,

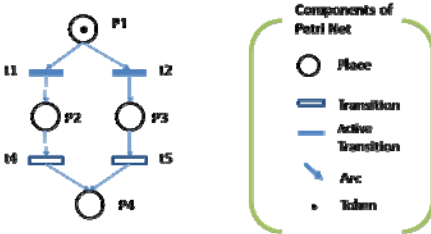


Fig. 1. Petri Net

or whether this is the user's first access, adding a sense of space and time to the digital library infrastructure. MIDAS, an extension to caT, adapts the information display to suit the characteristics of the client device [8]. For example, users accessing the hypertext from an iPhone might see different content than users of desktop computers.

Traditionally, authors have created caT nets using a graphical tool, called xTed, via point-and-click interaction to lay out net elements one-by-one on a two-dimensional canvas. The visual complexity of nets increases quickly as authors add more places and transitions to the net to include more content and to customize their behavior in response to different situations, such as a user's time of access, location, preferences, and privileges [4]. To help mitigate this problem, xTed supports authoring through hierarchical nets [6]. This approach enables authors to create a network of layered sub-nets and to deal with net segments without being overwhelmed by the larger network. Conceptually, this approach helps authors deal with structurally or functionally similar components together. Visually, a hierarchical net replaces parts of a network with a simpler visual surrogate for the sub-net that it represents. However, this support does not include the ability to combine or modify nets in a flexible manner. Typically, authors are overwhelmed when design networks with tens of nodes, making it impractical for them to design large nets that represent real-life situations in libraries and museums.

3 Template-Based Petri Net Composition

xTed supports only a bottom-up approach for creating nets. To author the small network shown in fig. 1, an author would create the place P2, followed by the transitions t1 and t2, before drawing the connecting arcs. The process quickly gets tedious to perform and boring to describe. To create this network, the author must think of the places, transitions, and their interconnections. However, in planning digital library services for their patrons, designers must take a top-down approach, thinking about high-level concepts to plan a network that meets their requirements.

Semantically, the net shown in figure 1 can be described as: the viewer, currently at place P1, has an option of viewing the contents either of place P2 or of place P3 by following the respective transitions, after which, she views the contents associated with place P4. Thinking of the network in behavioral terms enables authors to employ a top-down approach in designing purposeful caT networks. Our new interface, Template-based caT Authoring Tool, supports authors in selecting and modifying net components for efficient management and reuse by incorporating Petri net theory concepts, such as net transformation [1] and Petri net algebra [2]. Authors can organize net components into smaller, named units called templates.

3.1 Component Net and Template (Predefined Component Net)

To create nets by composing existing nets as building blocks, we employ a structure called the component-based Petri net. A component net (CN) is a net fragment that consists of a set of places, transitions, arcs, and subnets. Each component

net stores its characteristic and identifying meta data, such as name, description, functions, constraints, properties, summary, net type, media type, and structural pattern. For new nets the author provides metadata manually. When importing existing nets to construct a CN, the CN inherits some metadata from the parent nets, which helps identify components included in a large specification. Input and output ports (places) of a CN that connect to other components are defined in Single-Input/Single-output (SISO) form.

Our component nets support five composing operations from Petri net algebra to organize individual nets into larger structures: sequence ($;$), choice ($+$), parallel (\parallel), iteration (μ), and refinement (r) [2]. These operations enable authors to construct larger nets from existing fragments. For example, a digital museum curator who is putting together an online fine art exhibit may reuse existing collections of Vincent van Gogh's art (named VG) and Claude Monet's masterpieces (named CM) using various composition operations to achieve different effects. She could use the sequence operation to have her patrons first visit van Gogh and then Monet (VG;CM). Alternately, the choice operation provides patrons the option of browsing either collection (VG+CM). To enable her audience to compare the artists' styles, she may present the collections concurrently, using the parallel operation (VG \parallel CM). The other two composing operations affect single CNs. Iteration requires the browsing of the net fragment a certain number of times. Refinement replaces a transition with a subnet, akin to hierarchical net composition in xTed [6].

Templates are pre-defined CNs, which instantiate commonly used structures that aid users in customizing nets quickly. Templates have been used extensively for design tasks, both in the physical and digital realms. Several web sites offer free, downloadable templates for novice web designers and JavaScript programmers to accomplish common tasks [3][12]. Perhaps, the analogy that best describes caT templates is that of design patterns developed in the context of object-oriented programming [5]. caT templates provide a physical structure of places and transitions that embodies particular behavior by virtue of the associated browsing semantics. For example, a template for sequential viewing provides a set of linear organization of places to organize content. Viewers must negotiate this structure in the order laid out by the author, accessing the contents of one place at a time. Thus, the use of templates frees authors from the chore of putting together places, transitions, and arcs and enables them to focus on the interactive and behavioral features they wish to provide to the viewers of digital collections. Templates can easily be combined with the composition operations. Also, a template could be applied either to a single fragment or to each fragment of this form.

Sometimes, either by human oversight or as a result of the templating and composition operations, caT networks may develop redundant constructs. Transformations replace complex nets with functionally equivalent yet structurally simpler nets. We have implemented two transformations: union, which combines two structures with a shared substructure, and fusion, which combines two copies of a substructure within a structure [1]. For example, the reduction rule [11] removes parallel nodes; nodes that are structurally and functionally identical within a net. To be considered parallel, nodes must have identical input and output nodes, identical browsing semantics and the same information content associated with them. Conversely, parallel nodes can be added, when necessary, to support authoring goals.

3.2 Textual Authoring Language

Visual manipulation of net elements via point-and-click interaction is a slow process. Textual representation allows for compact viewing of large structures. It enables authors to recognize and locate named CNs easily within a large collection of net fragments. Our new authoring tool supports human-readable textual description to combine named templates, for example, $Seq(C1,C2)$, $Choice(C1,C2)$, $Parallel(C1,C2)$, $Iteration(C1)$ and $Refinement(C1,a, C2)$. While individual places and transitions can be created using the textual language, it is most useful when recomposing a net from pre-existing fragments.

3.3 Template-Based caT Authoring Tool

The Template-based caT Authoring Tool (TcaT) integrates the authoring improvements that we have described in this section in a familiar, desktop application interface, as shown in fig. 2. The interface displays the net structure in a drawing panel, in this case, Net ID 2 gives users the choice between browsing the information in NET ID 1 or NET ID 0, each of which display the contents of multiple places in parallel. Authors may create places, transitions, and connect these using arcs, much as they did with xTed. Instant views of the content for a place, author added metadata and annotations for net element are displayed as tool tips.

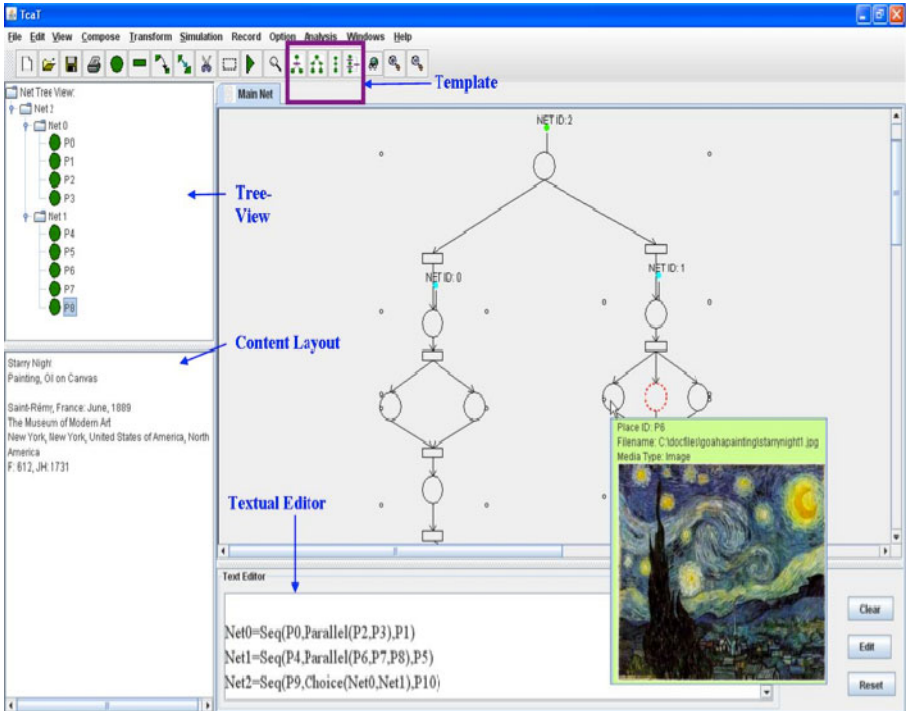


Fig. 2. Template-based caT Authoring Tool

Net collapsing simplifies net editing and management, while reducing the display complexity of large nets. Collapsed nets are represented by differently rendered place-sin the current net and remain available to authors for editing in different tabbed panels. When a collapsed net is expanded in the parent net, it replaces the surrogate place and the panel that displayed it is closed. For collapsed nets, tool tips display thumbnails of the original net structure for the selected node. This helps authors get a quick preview of the structure of collapsed nets.

The tree view presents the overall net structure to ease navigation among net components. A content layout window displays the content associated with a place on the canvas. TcAT includes a document editor for easy editing of place content. For editing complex document types that TcAT does not handle natively, double clicking the selected place invokes other applications such as Microsoft Office programs, web browsers, image viewers, and media players.

TcAT's text editor doubles as a textual authoring environment for caT component nets. TcAT converts the textual specification to the graphical form and vice versa automatically without any user intervention. Thus, the two forms are always synchronized with each other and an author may switch between the two seamlessly.

4 User Evaluation

We evaluated the effectiveness of TcAT with the help of 15 subjects, who have earned at least a bachelor's degree in various disciplines. We collected data by analyzing their performance during the authoring tasks. Based on their competencies, we classified the subjects into three groups: novice—well-versed in using computers and the Web, intermediate—computer science and engineering background, and expert—experience with graph theory and Petri nets. Our subject pool consisted of 5 novices, 6 intermediate users, and 4 experts. None of the subjects had used xTed or TcAT before their participation in this study. They received a brief training on using these tools before they beginning their tasks.

The subjects were to create a digital art museum using images of van Gogh's and Monet's art. We provided them with information about 17 paintings by van Gogh and 10 by Monet. The information included an image file, a summary that includes its title, category, date, exhibition location, and a few descriptive passages. Subjects could take as long as they needed to complete the task and were required to include all the information provided in order to create their galleries. Each subject performed the task using three techniques: xTed—the traditional authoring interface, TcAT with graphical interaction (TcAT-GI), and TcAT with textual language. The experts were familiar with Web page authoring and also created a Web site for the galleries using Microsoft FrontPage. We did not train them for this task.

The textual language emerged as the quickest method for this task. Every subject required more time to create their collection when using xTed than with either interface using TcAT. The experts required comparable time to create the collection using xTed and Web authoring. The subjects took the unlimited availability of time seriously. In spite of learning the use of caT authoring tools just prior to performing the tasks, subjects were quicker when using xTed as well as TcAT than they were with FrontPage (exception: Expert 3 with xTed).

The performance of the subjects for both TcAT methods—textual and graphical—is more consistent than with xTed. As expected, experts took the least time to complete the task, with the intermediate users close behind. The difference in authoring time is statistically significant between xTed and each of the TcAT modalities with $p < 0.05$, implying that TcAT supports the authors' tasks better than xTed. The difference in authoring time between TcAT-GI and the textual language is not statistically significant ($p = .207$). The difference in performance of novices vs. experts as well as that of novices vs. intermediate subjects is statistically significant ($p < 0.05$) but that between intermediates and experts is not ($p = .349$).

Table 1. Descriptive Statistics for Improvement Rate (%)

Comparison Pairs	Subject Type	Mean	Median	Std. Deviation	Range	Minimum	Maximum
xTed vs TcAT GI	Novice	42.50	43.14	9.84	26.85	28.85	55.70
	Intermediate	43.96	50.14	19.07	49.35	11.63	60.98
	Expert	55.37	60.49	14.23	31.52	34.48	66.00
	Total	46.51	48.00	15.23	54.37	11.63	66.00
xTed vs TcAT Language	Novice	45.77	47.06	15.98	44.01	23.08	67.09
	Intermediate	64.28	71.28	19.68	52.09	27.91	80.00
	Expert	59.13	61.84	10.27	23.17	44.83	68.00
	Total	56.73	58.82	17.40	56.92	23.08	80.00
TcAT GI vs Language	Novice	7.33	6.90	12.19	33.82	-8.11	25.71
	Intermediate	38.14	40.63	16.66	43.12	18.42	61.54
	Expert	7.20	6.51	6.52	15.79	0.00	15.79
	Total	19.62	15.79	19.89	69.65	-8.11	61.54

Table 1 presents the summary statistics of performance improvements for interface pairs. Our subjects' performance indicates that the improvement is the maximum between xTed and TcAT-textual language (56.73%) and the least between the two modalities of the TcAT (19.62%). The maximum and minimum difference in performance also follows the expected trend for each pair of interfaces. Analyzing the group-by-group statistics for the three interfaces, this table shows that the performance improvement between authoring tool pairs was consistent for all groups of users. Experts, intermediates, and novices, all demonstrated the most improvement between xTed and the TcAT textual language. This consistency of behavior across subject categories is a significant trend, which indicates that the new tools may be useful to all. This will help with improving the efficiency of those who have used xTed as well as with attracting new users to explore our tools.

5 Conclusion and Future Work

Digital libraries, museums, and collections employ several well-recognized, templatable features, such as on-line and off-line help, galleries, tutorials and self-help aids, creation and browsing of personalized collections, specialized or advanced services for paying patrons, and additional privileges for collection managers. caT provides a robust infrastructure for modeling these features as Petri net-based hypertext networks. TcAT expands the features provided by xTed, the traditional interface, to provide greater flexibility in authoring caT networks. Authors can create, edit, and save named templates and combine these using compositional mechanisms to model digital library behaviors using Component Net semantics. Synchronized visual-textual views let authors switch modalities at will to conveniently manipulate large structures.

In our pilot evaluation, novice as well as expert caT users authored a digital art gallery using TcAT significantly faster than their counterparts who used xTed. The performance improvement of TcAT users was slightly higher when using the textual authoring language. TcAT authors preferred the textual language and quickly adopted its use. Over time, we will continue to analyze the networks created by authors to assess the effectiveness of new features. We expect that the new features will encourage authors to create larger networks due to the ease of authoring.

Additional user studies will help us understand the cognitive process involved in authoring large caT networks that model complex interactions of patrons' use of digital libraries. Which features do authors design using top-down or bottom-up approaches? Which templates get used the most? Are there additional templates that authors would value for creating nets that support specific features? Future work also involves supporting additional composition operations to provide greater flexibility in repurposing networks to support new digital library features.

References

1. Berthelot, G.: Transformations and Decompositions of Nets. In: Brauer, W., Reisig, W., Rozenberg, G. (eds.) APN 1986. LNCS, vol. 254, pp. 359–376. Springer, Heidelberg (1987)
2. Best, E., Devillers, R., Koutny, M.: Petri Net Algebra. Springer, New York (2001)
3. Campbell, R.: Templates in Javascript, [http://particletree.com/notebook/templates-in-javascript/\(viewed on February 1, 2010\)](http://particletree.com/notebook/templates-in-javascript/(viewed%20on%20February%201,%202010))
4. Furuta, R., Na, J.-C.: Applying caT's Programmable Browsing Semantics to Specify World-Wide Web Documents that Reflect Place, Time, Reader, and Community. In: Proc. ACM Symposium on Document Engineering, McLean, VA, November 2002, pp. 10–17. ACM Press, New York (2002)
5. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison Wesley Professional, Upper Saddle River (1994)
6. Jensen, K.: Coloured Petri Nets - Basic Concepts. In: Analysis Methods and Practical Use. Basic Concepts. EATCS Monographs on Theoretical Computer Science, vol. 1, Springer, Berlin (1992)
7. Karadkar, U., Na, J.-C., Furuta, R.: Employing Smart Browsers to Support Flexible Information Presentation in Petri net-based Digital Libraries. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, pp. 324–337. Springer, Heidelberg (2002)
8. Karadkar, U., Furuta, R., Ustun, S., Park, Y., Na, J.-C., Gupta, V., Ciftci, T., Park, Y.: Display-agnostic Hypermedia. In: Proc. ACM Hypertext 2004, Santa Cruz, CA, August 2004, pp. 58–67. ACM Press, New York (2004)
9. Na, J.-C.: Context-aware Hypermedia in a Dynamically Changing Environment, Supported by a High-Level Petri Net. Ph.D. Dissertation, Texas A&M University (December 2001)
10. Stotts, P.D., Furuta, R.: Petri-net-based hypertext: Document structure with browsing semantics. ACM Transactions on Information Systems 7(1), 3–29 (1989)
11. Tu, S., Shatz, S.M., Murata, T.: Applying Petri net reduction to support Ada-tasking deadlock detection. In: Proc. of 10th International Conference on Distributed Computing Systems, Paris, France, pp. 96–103. IEEE-CS Press, Los Alamitos (1990)
12. Web search – Web page template, [http://www.google.com/search?q=web+page+template\(retrieved on February 1, 2010\)](http://www.google.com/search?q=web+page+template(retrieved%20on%20February%201,%202010))

Uncovering Hidden Qualities – Benefits of Quality Measures for Automatically Generated Metadata

Sascha Tönnies¹ and Wolf-Tilo Balke^{1,2}

¹ L3S Research Center, Appelstraße 9a, 30167 Hannover, Germany

² IFIS TU Braunschweig, Mühlenpfordstraße 23, 38106 Braunschweig, Germany
toennies@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. Today, digital libraries more and more have to rely on semantic techniques during the workflows of metadata generation, search and navigational access. But, due to the statistical and/or collaborative nature of such techniques, the underlying quality of automatically generated metadata is questionable. Since data quality is essential in digital libraries, we present a user study on one hand evaluating metrics for quality assessment, on the other hand evaluating their benefit for the individual user during interaction. To observe the interaction of domain experts in the sample field of chemistry, we transferred the abstract metrics' outcome for a sample semantic technique into three different kinds of visualizations and asked the experts to evaluate these visualizations first without, later augmented with the quality information. We show that the generated quality information is indeed not only essential for data quality assurance in the curation step of digital libraries, but will also be helpful for designing intuitive interaction interfaces for end-users.

Keywords: Digital Libraries, Information Quality, Semantic Technologies.

1 Introduction

Digital Libraries have to handle a vast amount of data ranging from individual papers or reports in journals, conference proceedings etc. up to complete digitized books. Making such data searchable relies mainly on the amount and quality of the provided metadata. On a purely bibliographic level this metadata is relatively easy to derive and maintain, in contrast on the content level the problem of deriving correct metadata obviously grows with the density of information. Whereas the information contained in short conference papers can be manually extracted and annotated quite easily, capturing the content contained in a book definitely needs automatic means of extraction. Today semantic techniques relying on statistically approaches like term co-occurrences or frequencies are already commonplace. But the quality of metadata derived by such techniques is largely uninvestigated. Thus a main topic of future digital library research has to be a quality assessment of such techniques. Obviously the quality – like in information retrievals' precision / recall analysis – can only be evaluated comparing the techniques output with manually provided judgments.

In our previous research about digital libraries [1] and large digital book collections [2] we proposed three general metrics, i.e. Degree of Category Coverage (DCC),

semantic word bandwidth (SWD) and relevance of covered terms (RCT), for measuring the quality of semantic techniques used for taxonomy / folksonomy creation. These quality measures were derived by observing the workflow of a domain expert using the example of (but not limited to) the field of chemistry. First evaluations already pointed out the frameworks' usefulness but a thorough investigation is still needed.

In this paper, we evaluated the metrics' usefulness taking the Semantic GrowBag technique as an example of automatic metadata generation techniques. Our contribution is threefold:

- First, we performed a user study with domain experts to assess the general applicability and usefulness of our quality measures in the field of chemistry.
- Second, we applied the quality measures to the Semantic GrowBag technique to demonstrate the added value of purely statistical techniques in metadata generation.
- Third, we showed that already simple diagram types are sufficient to transport the information provided by quality measures.

2 Related Work

Recently digital libraries have made the step towards using semantic techniques for *automated metadata generation*. Typical examples include for instance exploiting term co-occurrences or language models to find relevant keywords, categorization, or even inter-document relationships like for instance in JeromeDL [3]. But already in early projects the need to evaluate the quality of the metadata became clear, although it has usually been restricted to general user satisfaction studies.

An excellent overview of related work in the field of quality assessment of manually and automatically generated metadata and semantic annotation techniques is done in [1]. But of course the annotated information also has to be displayed to help users in efficient document retrieval and selection. Thus, up to a certain degree also *information visualization techniques* are related to our paper. The classical representations of taxonomies used for navigational access are acyclic directed graphs, usually trees. Also, the Semantic GrowBag technique [5] used in our evaluation generates graphs of the automatically generated taxonomies. On the other hand, light-weight ontologies or folksonomies are often simply represented by tag clouds. Due to the popularity of tag clouds a lot of work has already been done to investigate their possibilities for searching and browsing in large document collections. For instance, [6] tries to answer the question whether tag clouds provide sufficient value for information seeking. Tag clouds are characterized as particularly useful for browsing or non-specific information discovery. Moreover, tag clouds provide a compact visual summary of the content and scanning the tag cloud requires less cognitive load than formulating specific query terms. In contrast, building tag clouds requires a lot of effort, if the esthetic sensation of a user should be matched. Therefore, several approaches, e.g. [7], and [8] investigate the influence of text size and position as well as the algorithms to use for tag cloud generation. Still, all this work relies on the frequency of terms and not on the underlying quality.

3 Evaluating the Value of Quality Measures for Semantic Techniques

We conducted a user study with domain experts, in our case practitioners in the field of organic chemistry, for evaluating the three metrics DCC, SWD and RCT defined in [1]. The aim of the study was first to get a feeling whether the defined metrics are useful and second how important the information provided is compared to classical forms of visualization.

For the experiments we used a corpus of 4554 documents extracted from the Journal of Synthetic Organic Chemistry (SYNTHESIS) published by Thieme Publishers, Stuttgart, Germany. For all papers we extracted the author keywords (9554) and eliminated all those with little discriminating power (terms like ‘*experiment*’ or ‘*synthesis*’) occurring in many papers. For the remaining set of about 1600 terms folksonomies were generated using the Semantic GrowBag technique [5], which relies on higher order co-occurrences of keywords in relation to the respective documents. For the actual experiments we randomly chose ten query terms for each expert to evaluate the quality of the respective folksonomy. For each query term we generated three different kinds of visualization: the original GrowBag graph (Fig. 1, query term in black box), the respective Tag Cloud and a concentric circle diagram (CCD) (Fig. 2, query term in the center).

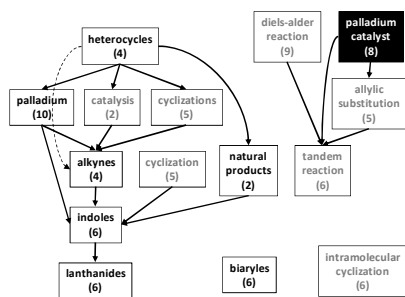


Fig. 1. The generated GrowBag graph for the keyword *palladium catalyst*

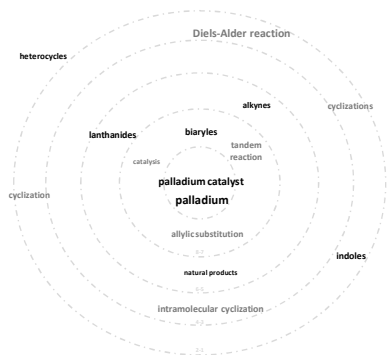


Fig. 2. The generated concentric circle diagram for the keyword

Basically the information provided by our three quality metrics, i.e. related category, overall specificity topical and distance to query term, can be represented by the visual features text color, text size and spatial layout. Please note, this information can be easily visualized in the CCD, whereas the other two visualizations lack the possibility to visualize the distance in an intuitive way. For the tag cloud we tried several clustering algorithms and visualized the terms in clusters, thus sacrificing the compactness of display for the possibility to show spatial relationships. However, since intra-cluster similarity usually clashed with the individual terms’ relationship to the query users tended to be confused by that notation. Hence, in our experiments we

tested the advantages of compact visualization versus the benefits of the information provided by the distances to the query term.

3.1 Designing the User Study Using the Semantic GrowBag Technique

To control the environment we ensured that all participating domain experts were recruited from the field of organic chemistry and in particular familiar with the focus area of the SYNTHESIS journal. Hence, we could expect only slight variations with respect to the individual knowledge spaces. Although the experts did not know about the specific semantic GrowBag technique used for deriving the graphs, all participants had prior experience with the use of ontological information retrieval and were proficient in using computing devices.

As stated above for all users we randomly selected ten query terms and confronted them with the three different visualizations in individual questionnaires. Filling in the questionnaire took only about half an hour of time. To illustrate the design of our user study let us focus on an example evaluation workflow for the query term '*palladium catalyst*'. The respective visual representations are shown in figures 1 and 2. Each questionnaire was divided into three major blocks:

- The first block of questions in the questionnaire focused on the first impression with respect to the diagrams. Users were asked to rank the different diagram forms for each query term regarding the intuitive understandability, i.e. the degree of ease to grasp the concepts contained.
- After evaluating the first impression the second block should prove if the users intuitively interpreted the diagrams in the correct way. Therefore, each metric and the correlation between the metrics' outcome and the diagrams were explained. With this knowledge the users were again asked to rank the different diagram forms.
- The third block actually measured the correctness of the three quality metrics. Therefore the domain experts were asked to rank the visualized metrics' outcome for each query term.

In more detail, the metrics explained in the second part of the evaluation concluded in the following visualization. Focusing again on the example query term '*palladium catalyst*', all terms can be categorized into the two categories, i.e. reactions (light) and chemical substances (dark). The size of each keyword represents the overall specificity, e.g. '*Diels-Alder reaction*' is quite specific term as it represents a concept of specific reaction with given reactants, products, solvents and reaction conditions. This way a domain expert reading the term 'Diels-Alder reaction' - maybe in connection with a substance - has a good impression of a reaction scenario and possible products. In contrast, the '*tandem reaction*' is an unspecific term describing a broader concept of a reaction type with much more space for interpretation. The term '*tandem reaction*' just defines a cascade of reactions from an educt to a product without the isolation of any intermediate product: the actual reactions of the cascade are not defined in detail by this concept. As already stated above, the closeness of a term in relation to the query term can only be visualized within the CCD, i.e. the distance of each keyword to the circles' center. Thus, the query term '*palladium catalyst*' is located in the circle center. Closely related terms like e.g. '*palladium*' and '*tandem reaction*' are located

nearby, whereas only loosely related terms are located far away e.g. ‘*heterocycles*’. The closeness of ‘*palladium catalyst*’ and ‘*palladium*’ is obvious as a palladium catalyst contains the metal palladium, which in turn defines the functionality of the catalyst. Tandem reactions are most often catalyzed reactions with a high stereo selectivity induced by various classes of palladium catalysts. An example for a loosely related term is ‘*heterocycles*’, which represents a general concept of a substance class with a rather weak relation to the term ‘*palladium catalyst*’.

For the last part of the experiment the domain experts were given a scale divided into five degrees (0 - 4) of satisfaction (see table 1).

Table 1. Evaluation scale for part 3 of the experiment

Value	DCC: percent of occurring concepts	SWD: percent of matching proportional font sizes	RCT: percent of matching distances
4 - completely satisfied	> 90%	> 90%	> 90%
3 - mostly satisfied	~ 75 %	~ 75 %	~ 75 %
2 - satisfied	~ 50%	~ 50%	~ 50%
1 - partially satisfied	~ 25%	~ 25%	~ 25%
0 - unsatisfied	≤ 10%	≤ 10%	≤ 10%

3.2 Experimental Results

In the first part of the experiment we evaluated the first impression and the intuitive understandability of the respective visualizations. We expected a high rank of the tag cloud as it is a compact and well known kind of visualization. Surprisingly, as can be seen in figure 3 the concentric circle diagrams (CCD) were already ranked considerably higher immediately claiming about 95% of the position one ranks with an average rank of 1.07. In contrast the tag cloud visualization just got an average rank of 2.1 and the remaining 5% of position one ranks. The (somewhat harder to understand) ontology graph was never ranked at position one and only got an average rank of 2.82.

It is interesting to note that in topically focused document collections quality information blended into navigational information or categories is indeed attractive for users. This also shows that even the simple CCD is already an intuitive way of visualization for our quality metrics. Also a later interview with selected domain experts confirmed this: they explained the lower rank of the tag cloud, because the co-occurring terms were sometimes misleading. But the adoption of the distance in the CCD clarified intuitively that some terms may belong to the query term in a rather loosely coupled way.

In the next step of the experiment, the visualizations’ semantics in terms of encoded quality measures was explained in detail and the domain experts were asked to re-rank the three kinds of visualization. As expected, the CCD was still most often ranked at position one (see figure 4). However, a marginal loss of 2.5% in position one ranks occurred, still resulting in 92.5% of top positions and an average rank of 1.1. At this stage, although gaining 7.5% of the overall position one ranks, tag clouds experienced a slight drop in the average rating (2.15). This can particularly be attributed to their

limitations becoming clear during the explanation of the semantics: users better understood their power of compact representation, but also their difficulties in discriminating terms. Again, the graph representation was never ranked first but the re-ranking still resulted in a slightly better average rating of 2.75. Further interviews with the domain experts have shown that they liked the tag cloud more than the CCD in situations, when confronted with very sparse CCD diagrams. This shows that there is a tradeoff between compactness of the visualization and the transported information. One possibility to handle this tradeoff would thus be a digital library interface where first a tag cloud of the digital collection is shown and once a user selects a term for deeper investigations, the respective CCD is shown.

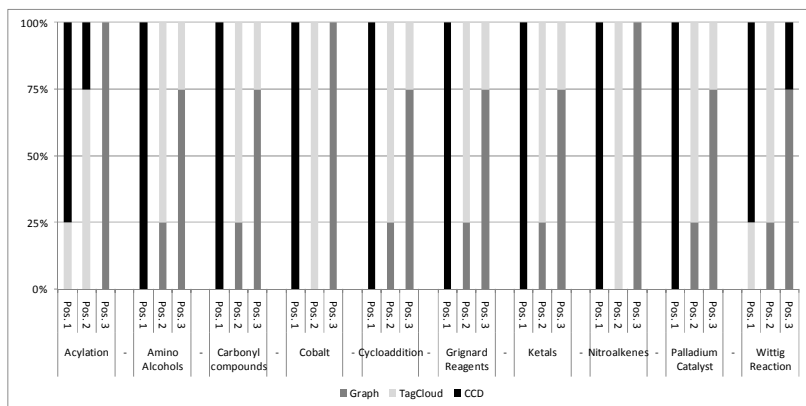


Fig. 3. First impression results

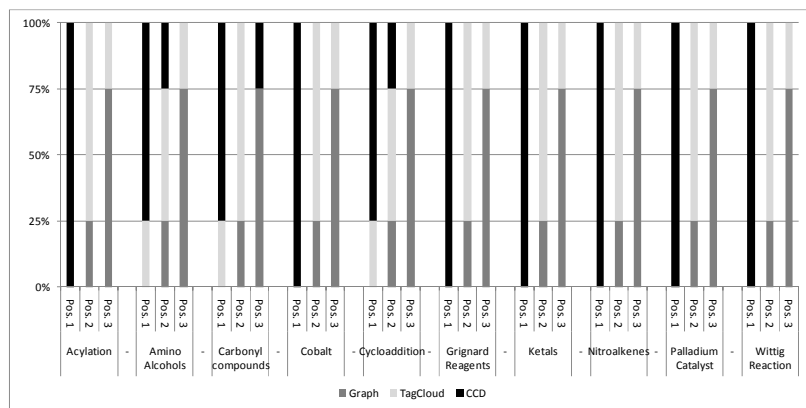


Fig. 4. Second impression results

Due to the fact, that the CCD is the only diagram which represents our metrics' entire outcome and that the rank has only slightly been decreased after explaining the quality metrics contained, our three metrics indeed seemed reasonable for the domain

experts. A deeper investigation as last part of the evaluation further substantiates this prediction. We asked all experts to consider the three metrics individually and evaluate the terms provided by the Semantic GrowBag technique for the ten test queries. As can be seen in figure 5 on a scale from 0 to 4 none of the metrics' outcome has been ranked less than 2, i.e. the 50% mark of satisfaction. In average the degree of domain coverage (DCC) was ranked with 3.20, the semantic bandwidth (SWD) with 2.82 and the relevance of covered terms (RCT) with 3.18. On average the domain experts were mostly satisfied with the quality of the Semantic GrowBag technique's generated metadata and also the proposed quality metrics' usefulness in quality assessment was confirmed.

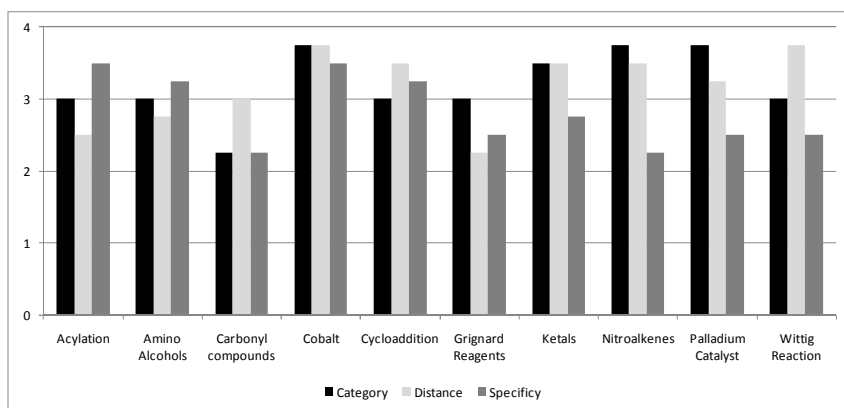


Fig. 5. Rating of the correctness of the quality aspects from unsatisfied (0) to completely satisfied (4)

4 Conclusions and Future Work

Since customers from academia and industry depend on timely and correct information provisioning, the focus of today's digital libraries has to be on information quality. Therefore, digital libraries to a large degree still rely on manual curation (in contrast to e.g., IR-based indexing in Web search engines). However, given the exponential growth of digitally available documents and the high costs of manual labor, also digital libraries soon will have to employ social or semantic techniques for automated metadata generation. But in order to still uphold the high quality standards, the quality of all metadata, and thus, ultimately the quality of the techniques used for metadata generation has to be assessed.

To address this problem, we proposed a set of three general purpose quality metrics for metadata generation by supervising the interaction of domain experts with metadata in the example field of chemical literature. In this paper we investigated the usefulness of the proposed metrics and their information gain. For this purpose, we conducted a user study with domain experts again relying on the field of chemistry as an example application. We showed that it is indeed useful to measure the quality of a semantic technique: the domain experts were easily able to assess the outcome of the

technique and gained insights into what quality to expect during their information gathering. The Semantic GrowBag, a statistic technique relying on term-co-occurrences for deriving metadata, was graded with an average of about ‘mostly satisfied’, i.e. about 75% complete, related, and specific. Moreover, by also providing the quality values visually for each term within the navigation elements, domain experts were less confused (especially when interacting with a low grade folksonomy).

For the investigation during our survey we used a very simple kind of diagram derived from fisheye view interfaces visualizing quality values of each term by its distance to the query term: the concentric circle diagram. Still, we were able to show that users are more satisfied by the experience of using this kind of diagram than the semantically rather shallow, yet popular tag clouds. In future work we also want to address the problem of different interfaces for visualizing quality information. The problem here seems to be twofold: one aspect is the uses of semantic techniques while indexing documents, which suggests a kind of ‘quality dashboard’ for the curator. The second aspect has to deal with the navigational elements used during user interactions for information seeking.

Acknowledgments. This work was supported by the German Research Foundation (DFG) within the program ‘Scientific Library Services and Information Systems’.

References

- [1] Tönnies, S., Balke, W.: Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 168–179. Springer, Heidelberg (2009)
- [2] Tönnies, S., Balke, W.: User-centered Content Provisioning over Large Collections of eBooks. In: Proceedings of the 2009 2nd ACM Workshop on Research Advances in Large Digital Book Repositories, Books Online 2009 Corfu, Greece, October 2 (2009)
- [3] Kruk, S.R., Woroniecki, T., Gzella, A.: JeromeDL – a Semantic Digital Library
- [4] Tönnies, S., Balke, W.: Using Semantic Technologies in Digital Libraries – A Roadmap to Quality Evaluation. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 168–179. Springer, Heidelberg (2009)
- [5] Diederich, J., Balke, W.: The Semantic GrowBag Algorithm: Automatically Deriving Categorization Systems. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 1–13. Springer, Heidelberg (2007)
- [6] Sinclair, J., Cardew-Hall, M.: The folksonomy tag cloud: when is it useful? *Journal of Information Science* 34, 15–29 (2007)
- [7] Rivadeneira, A.W., Gruen, D.M., Muller, M.J., Millen, D.R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: Conference on Human Factors in Computing Systems (2007)
- [8] Lohmann, S., Ziegler, J., Tetzlaff, L.: Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 392–404. Springer, Heidelberg (2009)

Query Transformation in a CIDOC CRM Based Cultural Metadata Integration Environment

Manolis Gergatsoulis^{1,*}, Lina Bountouri¹,
Panorea Gaitanou¹, and Christos Papatheodorou^{1,2}

¹ Dept. of Archives and Library Science, Ionian University, Corfu, Greece

² Institute for the Management of Information Systems (IMIS),
Athena R.C., Athens, Greece

{manolis,boudouri,rgaitanou,papatheodor}@ionio.gr

Abstract. The wide use of a number of cultural heritage metadata schemas imposes the development of new interoperability techniques that facilitate unified access to cultural resources. In this paper, we focus on the ontology based semantic integration by proposing an expressive mapping language for the specification of the mappings between the XML-based metadata schemas and the CIDOC CRM ontology. We also present an algorithm for the transformation of XPath queries posed on XML-based metadata into equivalent queries on the CIDOC CRM ontology.

Keywords: Metadata interoperability, semantic integration, query transformation, mapping languages, metadata schemas.

1 Introduction

Cultural heritage institutions, archives, libraries, and museums host and develop various collections with heterogeneous types of material, often described by different metadata schemas. Managing metadata as an integrated set of objects is vital for information retrieval and (meta)data exchange. To achieve this, interoperability techniques have been applied [2]. One of the widely implemented techniques is the Ontology-Based Integration. Ontologies provide formal specifications of a domain's concepts and their interrelations and act as a *mediated schema* between heterogeneous sources [7].

This paper is motivated by a real life integration scenario of a set of data sources each of them providing information encoded by a different metadata schema (e.g, EAD, VRA, DC, MODS, etc.). Each schema is semantically mapped to the CIDOC CRM [3] based mediator, which may also retain its own database of metadata encoded in CIDOC CRM. Users can pose their queries either to the local data sources (following the format of the queries specified by these sources) or to the mediator. The local query engine (resp. the mediator's query engine) returns the results from its own database and promotes the query to the

* Part of this work was done while the author was on sabbatical leave at the Institute for the Management of Information Systems (IMIS), Athena R.C., Athens, Greece.

mediator. The mediator, based on the defined mappings, translates the query to suitable forms and forwards it to be answered by the other local sources. Finally, all results are collected and returned to the user.

This paper builds upon the mappings of DC, EAD and VRA metadata schemas to CIDOC CRM developed in [9,4] and defines the Mapping Description Language (MDL), to formally describe the semantic mappings between XML-based metadata schemas and CIDOC CRM. Besides, an algorithm that transforms XPath to RQL-like queries applied to the CIDOC CRM is presented. A real world example of the MDL mapping rules and the query transformation is demonstrated through the mapping of EAD [8] to CIDOC CRM.

2 The Mapping Description Language (MDL)

2.1 Introducing CIDOC CRM

The *CIDOC Conceptual Reference Model* (CIDOC CRM) is a formal ontology consisting of a hierarchy of 86 classes and 137 properties. A *class* (also called *entity*), identified by a number preceded by the letter “E” (e.g. E7 Activity, E31 Document), groups items (called *class instances*) that share common characteristics. A class may be the *domain* or the *range* of *properties*, which are binary relations between classes. Properties are identified by numbers preceded by the letter “P” (e.g. P14 carried out by (performed) with domain the class E7 Activity and range E39 Actor, P102 has title (is title of) with domain E71 Man-Made Thing and range E35 Title and P108 has produced (was produced by) with domain E12 Production and range E24 Physical Man-Made Thing). An *instance of a property* is a relation between an instance of its domain and an instance of its range. A property can be interpreted in both directions (active and passive voice), with two distinct but related interpretations. A *subclass* is a class that specializes another class (its *superclass*). A class may have one or more immediate superclasses. When a class *A* is a subclass of a class *B* then all instances of *A* are also instances of *B*. A subclass inherits the properties of its superclasses without exception (*strict inheritance*) in addition to having none, one or more properties of its own. A *subproperty* is a property that specializes another property. If *P* is a subproperty of a property *Q* then 1) all instances of *P* are also instances of *Q*, 2) the domain of *P* is either the same or a subclass of the domain of *Q*, and 3) the range of *P* is either the same or a subclass of the range of *Q*. Some properties are associated with an additional property (called *property of property*) whose domain contains the property instances and whose range is the class E55 Type. Properties of properties are used to specialize their parent properties.

2.2 MDL Language Definition

A mapping from a source schema to a target schema transforms each instance of the former into a valid instance of the latter. The proposed mapping method between the metadata schemas and CIDOC CRM is based on a path-oriented approach; hence the metadata paths are mapped to semantically equivalent CIDOC

CRM paths. Since the metadata are XML-based, the source paths are *location paths* of XPath [11] enriched with *variables* and *stars*. A mapping rule consists of two parts: the left one represents a path in the XML document and the right the corresponding CIDOC CRM path. Variables are used in both parts to declare and refer to branching points, while stars declare transfer of value from the XML element/attribute to the corresponding class instance. The syntax of the MDL *mapping rules* is given bellow in EBNF:

```

R ::= Left ‘--’ Right
Left ::= APath | VPath
APath ::= ε | ‘/’ RPath
RPath ::= L | L ‘*’ | L ‘{’ Vl ‘}’ | L ‘*’ ‘{’ Vl ‘}’
VPath ::= ‘$’ Vl ‘/’ RPath | ‘$’ Vl ‘{’ Vl ‘}’
Right ::= Ee | Ee ‘→’ O | ‘$’ Vc ‘→’ O | ‘$’ Vp ‘→’ ‘E55’
O ::= Pe ‘→’ Ee
O ::= O ‘→’ O
Ee ::= E | E ‘{’ Vc ‘}’ | E ‘{’=’ String ‘}’
Pe ::= P | P ‘{’ Vp ‘}’

```

where L represents relative location paths (of XPath), E (resp. P) represents class (resp. property) ids of CIDOC CRM, V_l *location variables*, V_p *property variables* and V_c *class variables*. Note that a) property/class variables appear either enclosed in curly brackets (representing *property/class variable definition*) or prefixed by ‘\$’ (representing *property/class variable consumption*); b) in the CIDOC CRM paths class/property ids are used instead of their full names.

3 Mapping EAD to CIDOC CRM

Archival description represents an *archive*, which is a complex set of materials sharing common provenance, regardless of form or medium. *Encoded Archival Description* (EAD) is widely used to create *electronic finding aids*, which materialize the archival description and its hierarchical structure, beginning from the whole archive and proceeding with its sub-components, the sub-components of sub-components, and so on. An *EAD document*, starting from the ead root element, consists of three basic elements: the mandatory *EAD Header* (eadheader), the optional *Front Matter* (frontmatter), and the mandatory *Archival Description* (archdesc). The latter carries the archival description itself providing identification (such as the archive’s title (unittitle) and creator (origination)), administrative and supplemental information, and the description of the hierarchical archival components (dsc). An archival component is a recognizable entity, characterized by an attribute level as *series*, *subseries*, *file*, *item* etc. Components are deployed as nested elements, called either c or c01 to c12.

Example 1. In this example we present a fragment of an EAD document:

```

<ead>
<eadheader>...</eadheader>
<archdesc level="fonds">
<did>
<unitid countrycode="GR" repositorycode="IU">ARC.14</unitid>

```

```

<unittitle>Ionian University Archive</unittitle>
<unitdate>1984 - 2007</unitdate>
<origination><corpname>Ionian University</corpname></origination>
</did>
<bioghist><p>The Ionian University was founded in 1984...</p></bioghist>
<controlaccess><corpname>Ionian University</corpname></controlaccess>
<dsc>
<c01 level="series">
<did>
<unitid countrycode="GR" repositorycode="IU">ARC.14/1</unitid>
<unittitle>R. C. Archives</unittitle>
<unitdate>1998 - 2007</unitdate>
<origination><corpname>I.U.Research Committee</corpname></origination>
</did>
</c01>
<c01 level="..."> ... </c01>
</dsc>
</archdesc>
</ead>

```

MDL can be used to describe the mapping of EAD to CIDOC CRM. Part of this mapping, containing the rules that map most of Example [II](#), are shown in Table [II](#). In the rest of this section, a short analysis of the MDL rules' semantics is presented: For example, rule *R1* states that the EAD document is mapped to the E31 Document class, i.e. it is an instance of this class. *R2* maps the archdesc subelement to a concatenation of CIDOC CRM *class - property - class*. In detail,

Table 1. Mapping rules: Mapping EAD to CIDOC CRM

R1:	/ead{X0}	E31{D0}
R2:	\$X0/archdesc{X2}	\$D0→P106→E31{D2}→P70→E22{A0}→ P128→E73{I0}
R3:	\$X2/@level*{Y2}	\$A0→P2→E55{A01}
R4:	\$Y2	\$A01→P71→E32 {=level}
R5:	\$X2/did/unitid*	\$A0→P1→E42
R6:	\$X2/did/unittitle*	\$I0→P102→E35→P1→E41
R7:	\$X2/did/origination{X22}	\$A0→P108b→E12{A03}
R8:	\$X22/corpname*	\$A03→P14→E40→P1→E41
R9:	\$X2/controlaccess/corpname*	\$I0→P67→E40→P1→E41
R10:	\$X2/dsc/c01{X24}	\$D2→P106→E31{D3}→P70→E22{A1}→ P128→E73{I1}
R11:	\$X24/@level*{Z2}	\$A1→P2→E55{A11}
R12:	\$Z2	\$A11→P71→E32 {=level}
R13:	\$X24/did/unitid*	\$A1→P1→E42
R14:	\$X24/did/origination{X242}	\$A1→P108b→E12 {A13}
R15:	\$X242/corpname*	\$A13→P14→E40→P1→E41
R16:	\$X24/did/unittitle*	\$I1→P102→E35→P1→E41

R2 states that i) archdesc corresponds to an instance of E31 Document, which is linked, through the binary relation P106 is composed of, to the instance of E31 Document representing the ead element, ii) the instance of E31 Document documents (P70 documents) an instance of E22 Man-Made Object (corresponding to the archive itself), iii) the archive (the instance of E22 Man-Made Object) carries (P128 carries) information, which is an instance of the class E73 Information Object. The instances of E31, E22 and E73 represent respectively a) an immaterial item that makes propositions about the reality, b) the archive as a physical object created by human activity, and c) the archive as information carrier.

The application of the above rules to the EAD document of Example 1, results to the CIDOC CRM data graph depicted in Figure 1. In this graph the upper part of a box indicates the EAD path mapped to the CIDOC CRM class instance shown in the lower part. Each instance has an instance id, denoted by an “o” followed by an integer. Specific instance values appear as assignments in the lower boxes. Boxes are linked with arrows representing CIDOC CRM properties. More detailed analysis of MDL’s semantics can be found in [4].

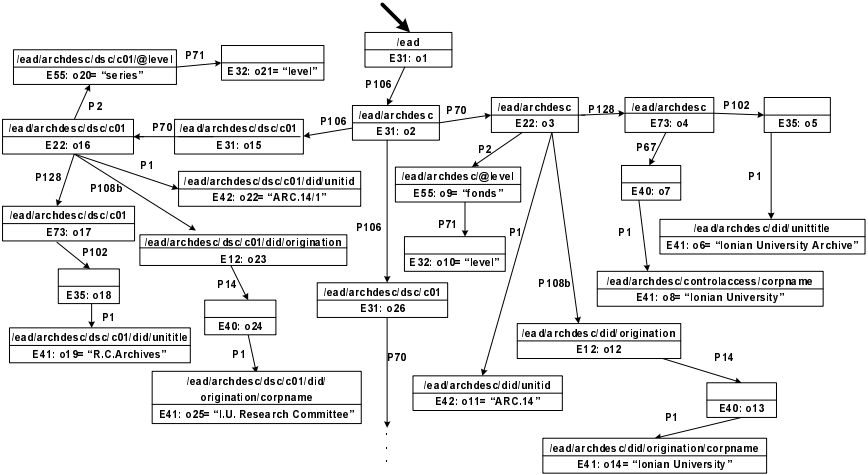


Fig. 1. The CIDOC CRM representation of the EAD document of Example 1

4 XPath to CIDOC CRM Query Transformation

In this section we present an algorithm, which uses MDL rules for transforming XPath queries posed on XML-based metadata to equivalent queries on CIDOC CRM ontology, written in an RQL-like syntax [10]. We consider XPath queries allowing only the *child axis* and predicates. The RQL-like syntax allows queries in the form of *select-from-where* clauses. The requested variables are inserted in the *select* clause. The *from* clause contains data paths expressed as triples, denoting a property and its domain and range, associated with data variables. The reuse of a specific variable in more than one data path expressions introduces joins between the triples. Finally, the *where* clause is used for data filtering.

Phase 1 of the algorithm: The input of Phase 1 consists of an XPath query \mathcal{X} and the mapping rules \mathcal{M} . The output is a sequence \mathcal{S} of *generalized CIDOC CRM data path expressions* (GDPEs) of the form:

$$CE_0 \rightarrow P_1 \rightarrow CE_1 \rightarrow P_2 \rightarrow \dots \rightarrow P_n \rightarrow CE_n$$

where P_i , with $1 \leq i \leq n$, are properties, while CE_j , with $0 \leq j \leq n$, are *class expressions*. A class expression CE_j consists of a class id E followed by an optional class variable definition (i.e. a class expression is of the form $E\{V_c\}$). CE_0 may also be of the form $\$V_c$, where V_c is a class variable, while CE_n may be followed by a ‘+’ sign and/or a predicate. Class variable definitions introduce linking points between GDPEs, while class variables prefixed by ‘\$’ refer to these linking points. The symbol ‘+’ marks the class whose instance corresponds to the query result. Finally, predicates use comparison operators ($=$, $>$, \geq , $<$, \leq) to impose constraints (inherited from the predicates of the XPath query) on the values of class instances. To construct \mathcal{S} , Phase 1 traverses \mathcal{X} *depth-first-left-to-right* and decomposes it into smaller paths matching the left parts of rules in \mathcal{M} . Based on the right part of the rules, the corresponding GDPEs are constructed. A crucial point is to keep track of the class variable definitions and to associate them with appropriately chosen data variables. Given that a rule may be used multiple times, a *variables’ renaming mechanism* is employed to provide appropriate names to the class variables of the right hand side of this rule, ensuring the consumption of the correct version of them.

Example 2. Consider the query: *Find the title of the archive identified as “ARC.14”*, on the EAD document of Example 1. This query can be expressed as an XPath:

/ead/archdesc/did[unitid=“ARC.14”]/unittitle

Applying Phase 1 of the algorithm to this query, we get the following GDPEs:

/ead	E31{D0}	(Rule R1)
/archdesc	\$D0→P106→E31{D2}→P70→E22{A0}→ P128→E73{I0}	(Rule R2)
/did/unittitle	\$I0→P102→E35→P1→E41+	(Rule R6)
/did/unitid=“ARC.14”	\$A0→P1→E42=‘ARC.14’	(Rule R5)

Notice in Figure 2 that the XPath has two branches rooted at the did element on which we apply R5 and R6 respectively.

Phase 2 of the algorithm: The GDPEs obtained in Phase 1 are transformed in RQL-like form. To construct the **from**-part, a fresh data variable is associated with each class appearing in GDPEs and triples of the form $\{X1; E1\}P\{X2; E2\}$, corresponding to GDPEs’ sub-paths, are inserted, where the pairs $\{X1; E1\}$ and $\{X2; E2\}$ describe the domain and the range of the property P . The variable related to the class marked with ‘+’ appears in the **select**-part, while the constraints appearing in GDPEs are used to construct the **where**-part of the query.

Example 3. (Continued from Example 2) Applying Phase 2 of the algorithm to the GDPEs obtained in Example 2, we get the following RQL-like query 1.

¹ For clarity reasons, we use here the full names of CIDOC CRM classes and properties.

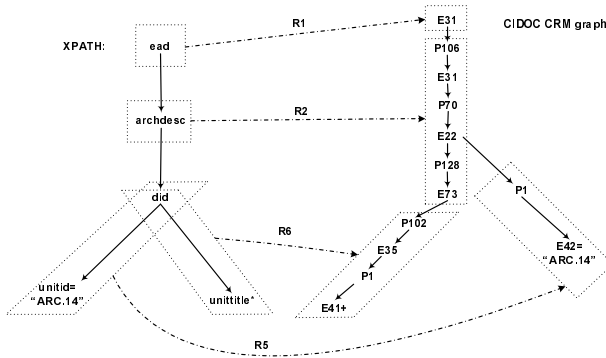


Fig. 2. Phase 1 of the transformation of the query in Example 2

```

select X6
from
{X1;E31_Document}P106_is_composed_of{X2;E31_Document},
{X2;E31_Document}P70_documents{X3;E22_Man-Made_Object},
{X3;E22_Man-Made_Object}P128_carries{X4;E73_Information_Object},
{X4;E73_Information_Object}P102_has_title{X5;E35_Title},
{X5;E35_Title}P1_is_identified_by{X6;E41_Appellation},
{X3;E22_Man-Made_Object}P1_is_identified_by{X7;E42_Identifier}
where X7=' ARC.14'

```

More complex XPath queries: So far we assumed XPath queries free of wildcards and descendant edges. However, there are significant queries, which make use of these constructs, such as the query “Find all the corporate names related to the archive”, expressed in XPath as `/ead//corpname`. A naive approach to treat such queries is to transform them into a set of XPath queries by eliminating the descendant axis (`//`) using the XML Schema or DTD specification of the metadata schema. Then, from the above query, we get the following XPath queries:

- (i) `/ead/archdesc/controlaccess/corpname`
- (ii) `/ead/archdesc/dsc/c01/did/origination/corpname`
- (iii) `/ead/archdesc/did/origination/corpname`

By applying the proposed algorithm we find an RQL-like query for each one of these queries. A similar approach can be used for queries containing wildcards.

5 Discussion

Recently, significant research efforts are focusing on Ontology-based Integration methods. In [1] queries are formulated in terms of the global schema (resembling CIDOC CRM) and then are translated in terms of the local XML sources. This architecture is not directly comparable with ours, since in our approach queries may be posed on the local sources (whose specification might be more familiar to the users) and then are reformulated in terms of the ontology. Nevertheless, we

mention this approach, since a path-to-path mapping language is defined to map the XML local sources to the ontology. Although this mapping language presents various differences with MDL, both languages share some common ideas. Among the novel characteristics of MDL is its capability of representing the “property of property” constructs, and the definition and use of (multiple) variables in both sides of the rules to declare and refer to linking points in the XML or CIDOC CRM paths. Besides, it allows predicates in XPath queries. The BRICKS Project [6] is a P2P architecture for cultural heritage institutions. Part of this architecture is the Metadata Manager tool, which provides various mapping options of the institutions’ local sources to the integrated system, such as mapping to CIDOC CRM. In BRICKS no generic mapping specification has been provided and mappings are implemented through spreadsheets.

The techniques presented in this paper are capable of transforming widely used metadata, such as VRA, MODS, and EAD to CIDOC CRM, taking into account the mappings defined in MDL by cultural heritage specialists. Based on these mappings, query rewriting is promoted, enhancing semantic integration.

References

1. Amann, B., Beeri, C., Fundulaki, I., Scholl, M.: Ontology-Based Integration of XML Web Resources. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 117–131. Springer, Heidelberg (2002)
2. Chan, L.M., Zeng, M.L.: Metadata interoperability and standardization - A study of methodology, Part I: Achieving interoperability at the schema level. *D-Lib Magazine* 12(6) (2006)
3. CIDOC CRM Special Interest Group. Definition of the CIDOC Conceptual Reference Model. Technical report (January 2010)
4. Gergatsoulis, M., Bountouri, L., Gaitanou, P., Papatheodorou, C.: Mapping Cultural Metadata Schemas to CIDOC Conceptual Reference Model. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS (LNAI), vol. 6040, pp. 321–326. Springer, Heidelberg (2010)
5. Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., Scholl, M.: RQL: A Declarative Query Language for RDF. In: WWW 2002, pp. 592–603. ACM Press, New York (2002)
6. Meghini, C., Risse, T.: BRICKS: A Digital Library Management System for Cultural Heritage. *ERICIM News* (61) (April 2005)
7. Noy, N.F.: Semantic Integration: a Survey of Ontology-Based Approaches. *SIGMOD Record* 33(4), 65–70 (2004)
8. Library of Congress (LC). EAD 2002 Schema (2002), <http://www.loc.gov/ead/eadschema.html>
9. Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Ontology-Based Metadata Integration in the Cultural Heritage Domain. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 165–175. Springer, Heidelberg (2007)
10. Theodoridou, M., Tzitzikas, Y., Doerr, M., Marketakis, Y., Melessanakis, V.: Modeling and Querying Provenance using CIDOC CRM. Technical Report Draft 0.94, Institute of Computer Science, FORTH-ICS (December 2008)
11. W3C CONSORTIUM. XML Path Language (XPath) 2.0. (January 2007), <http://www.w3.org/TR/xpath20/>

User-Contributed Descriptive Metadata for Libraries and Cultural Institutions

Michael A. Zarro and Robert B. Allen

College of Information Science and Technology, Drexel University
{michael.a.zarro,rba}@drexel.edu

Abstract. The Library of Congress and other cultural institutions are collecting highly informative user-contributed metadata as comments and notes expressing historical and factual information not previously identified with a resource. In this observational study we find a number of valuable annotations added to sets of images posted by the Library of Congress on the Flickr Commons. We propose a classification scheme to manage contributions and mitigate information overload issues. Implications for information retrieval and search are discussed. Additionally, the limits of a “collection” are becoming blurred as connections are being built via hyperlinks to related resources outside of the library collection, such as Wikipedia and locally relevant websites. Ideas are suggested for future projects, including interface design and institutional use of user-contributed information.

Keywords: Annotation, Descriptors, Metadata, Social Media.

1 Introduction

Contributions of metadata by users to online library holdings is an emerging phenomenon where the control of metadata generation is shared between librarians and curators, the traditional gatekeepers, and patrons. Notably, the Flickr Commons (flickr.com/commons) on the photo-sharing website Flickr.com allows cultural institutions to post digitized resources for public tagging, comment, annotation, and discussion. The implications of this phenomenon are significant in that patrons will have increased information and description of a resource, an enhanced ability to find resources through search and browsing, and use of the resource as a connector to additional materials on the Web.

The public’s response to the efforts of the United States Library of Congress, the first participating institution, has been “overwhelming” [8] and the resulting set of user-contributed metadata is a valuable source of descriptive information that may be utilized for information retrieval, resource identification, and outreach. The public has also taken it upon themselves to create linked, or “extended collections” connecting an institution’s resources to other material on the web; such as Wikipedia articles, commercial web pages and personal Flickr images. The use of Web 2.0 technologies and “social-sharing” websites has allowed libraries and other institutions a way to capture the collective intelligence of their patrons.

Comments and notes are types of “descriptive metadata” that describe, identify, and add context to a resource. User-contributed metadata enhances the identification of resources, connects resources to similar holdings in a private collection, and supports the retrieval of a resource through search and browsing. The ability to annotate a cultural institution’s resources is a significantly different experience from tagging, and one that has not yet been fully explored.

Given the lack of resources for cataloging and indexing historically available in libraries, patrons may be able to give a level of attention to rapidly growing digital collections that library staff cannot be expected to provide. With high processing speeds and bandwidth, users are equipped with systems fully capable of working with virtual image collections [10]. It is important to note that library patrons here are considered those who access library holdings wherever they exist. Thus, a visitor to the photo sharing website Flickr viewing a digitized photograph posted by a library is considered a patron of that library. Also, included in this definition are visitors to the library website where user contributions may be collected and/or displayed.

The Library of Congress in January of 2008 first posted photos to The Commons on Flickr (www.flickr.com/commons). The Commons is an area of the site where museums, libraries and other cultural institutions may post digital images for interaction with the community. The ongoing Library of Congress pilot project has three main objectives in sharing their content on the Flickr Commons; to increase awareness of the photographs in the Library’s holdings with people not likely to visit the Library’s website, gain experience with Web communities, and to understand how user-contributions may enhance the Library’s holding [8]. The Library invited users to annotate the images and was rewarded with a wealth of historical information and individual contributions.



Fig. 1. Typical resource from the Photochrom travel set posted by the Library of Congress on the Flickr Commons

Much previous research in the area of user-contributed metadata has studied users tagging their own resources, rather than those held by others or institutions. Marshall [7] described the tagging and annotation contributions for several hundred images of a popular mosaic in Venice and suggested that people may be better at story telling through titles and narrative metadata than they are at assigning one word labels and tags. An analysis of user contributions and user queries in the National Archives of the Netherlands found that the most popular query type is for geographic locations, while the most popular comment type is for named persons, places, or objects [10]. Thus, “stories” about named entities may prove to be most useful.

We aim to expand understanding of this phenomenon in two areas. First, we explore the annotation of a library’s holdings in the Flickr Commons by the public at large using three sets of digitized images that elicited significant user interaction. Second, we analyze comments and notes contributed by the user for the purposes of:

1. Expansion of historical or factual knowledge about the holding through user-contributed statements of fact or personal recollection.
2. Linking Out [2], or linking the holding to other resources such as personal Flickr photos or Wikipedia articles, through hyperlinks or textual pointers. In some cases, multiple links are contributed for a single library resource, building an “extended collection” related to a subject in the original, or “seed” image.
3. Corrections to existing descriptions and translations, as supplied by the Library or the public.
4. Linking In [2], or adding the Library’s images to groups of similar images.

2 Data Collection

We collected user-contributed metadata in the form of comments and notes for three images sets containing 1043 total images hosted by the Library of Congress on Flickr Commons. We downloaded as a tab delineated file the Flickr identifier, image title, and digest of tags for each image in three unique sets. First, we analyzed the Photochrom Travel Views set held by the Library of Congress. This set contains digitizations of color photochroms for various locations in Scandinavia, the British Isles, and Canada created between 1890-1900 [3]. There were a total of 657 images in the set at the time of data collection in October and November, 2009. Next, we looked at a digitized set of 364 black and white illustrated newspaper supplements published beginning in the 1880’s. The final set we examined was labeled “Mystery Pictures.” It is a somewhat smaller group of 22 color images, depicting content similar to the Photochrom set. In this case, the Library posted a call to action on the set homepage,

“HELP!! Please let us know if you recognize any of these images. France and the Mediterranean coast are likely locations for these travel views, called photochroms, from ca. 1900. They are the only pictures in a collection of 6,000 that arrived without titles.”¹

We manually examined each resource’s notes and comment stream to determine the types of metadata present. The categorization of comments was non-exclusive.

¹ http://www.flickr.com/photos/library_of_congress/sets/72157623063035332/

Because these can be paragraph length pieces of text, a single comment could be judged to belong to more than category (fig. 2). Multiple comments in a category were not tabulated, only the presence of at least one comment of the type. The intent was to determine at a high level, the presence of valuable contributions across the set.

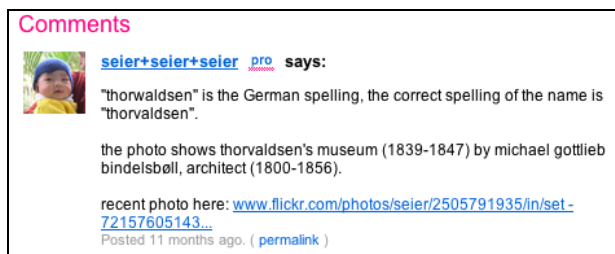


Fig. 2. Comment for resource shown in Figure 1. This single comment includes a correction, historical fact, and connector to a more recent photo of the subject.

Only English language comments were used in this study. Given the international flavor of many social sharing sites and the content, with users and images from around the globe, this may prove to be a shortcoming to be addressed in future studies.

3 Data Set

We found 37% of resources in the Photochrom set had at least one comment judged to fall within at least one of our categorizations listed above. As shown in figure 2, a comment may contain multiple categories of comments in just a few sentences. For example, a comment indicating a building was renovated in the 1950's that also contained a link to a photo of the reconstructed building counted as both a link out comment and a personal/historical comment. The counts shown in Table 1 indicate images where at least one contribution in the category was judged to appear.

The sets with geographic focus (Photochrom and Mystery) have a higher percentage of user-contribution than found for the Illustrated Newspaper Supplements, which received a contribution on 25% of its 364 images. Additionally, the set where the Library specifically asked for help – Mystery Photos - 100% of the images received a contribution. The small size of this set, 22 images, versus 657 for the Photochroms and 364 in the Newspaper Supplements, may be a factor.

These user-contributions are indexed by commercial search engines today and are being used to return results in web searches. For example, the phrase "Consul F.G. Gade" appears only in the comments of an image of Fantoft Church, in Bergen, Norway². A search in early December, 2009 for the term "Consul F.G. Gade" returned results in both Google and Bing. This term appears only in a user-contributed comment, it was not a tag or other form of metadata at the time of the search. Using Bing, the Flickr resource was the first result with and without the use of quotes designating a search for the exact phrase. On Google, using quotes to search for the exact phrase,

² http://www.flickr.com/photos/library_of_congress/3175010584/

the Flickr page is the first result. It is noteworthy that the search term does return a match in the local Flickr site search.

Table 1. Metadata categorization and results

Set	Metadata Category	Images	Percent
Photochrom Views	Personal and Historical	165	25%
	Link Out	105	16%
	Correction and Translation	83	13%
	Link In	86	13%
Illustrated Newspaper Supplements	Personal and Historical	62	17%
	Link Out	48	13%
	Correction and Translation	5	1%
	Link In	16	4%
Mystery Pictures	Personal and Historical	21	95%
	Link Out	16	73%
	Correction and Translation	4	1%
	Link In	0	0%

Corrections of cataloging mistakes in the library record by a user have the potential to be a great benefit of a commenting system. For example, the comment by chis-tenm.hielberg, “The correct spelling is Tyssestrengene (meaning the Tysse Strings, where Tysse refers to the catchment area). These beautiful "strings" (about 300 meters free fall) are part of a large hydroelectric power scheme and are sadly no longer to be seen” on a page titled “Tâysssestrengene, Hardanger Fjord, Norway.”³ A search on the translation of the corrected term, “Tysse Strings” is the first result in both Google and Bing web searches.

In these two examples, we see user-contributions being used to enable the finding of library resources. It is unlikely that cataloging or indexing staff in a typical library in the United States would have access to this level of local knowledge for a foreign resource. While they may certainly find this knowledge through research, the access to a crowd-based solution may prove to be more effective and more efficient.

4 Discussion

The “folksonomic flaw” as described in [4], suggests that tags may not be very useful for public information retrieval. Tags attached to a resource are generally for the use

³ http://www.flickr.com/photos/library_of_congress/3174183473/

of the submitter and have a personal meaning. There is little control over the submission of tags, and meanings can be ambiguous or hidden to the reader. For example, the tag “spike” may have many meanings as a noun or verb. The user searching for an image likely has little or no concept of the context in which the tag was submitted, and therefore may be presented many results in a search that are in fact not relevant in context. In contrast, the comments and notes we studied appear to be meant almost entirely intended to add context to an image.

Today, there are no widespread accepted practices for identifying the nature of user-contributions. Several categorizations have been created for metadata, particularly tags, in a social-sharing environment. [6] [7] [8] [10]. To make sense of the contributions we examined, they were grouped into four categories; Personal and Historical, Link Out, Corrections and Translations, and Link In. Contributions classified as Personal and Historical are related to a known historical fact or a user’s personal knowledge of the subject of an image. An example of a historical fact might be an architect’s name for a building in the photo. An example of a personal fact is shown in Figure 3.

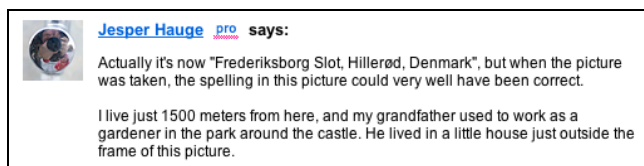


Fig. 3. A personal recollection submitted as a comment

While it may seem trivial to include personal facts, from our example we would learn that the castle mentioned had a garden of sufficient size and/or complexity as to require the employment of at least one groundskeeper. We hope the inclusion of personal remembrances give the researcher access to “hidden” facts about a resource.

The Link Out comments are connections from the current resource to additional resources on the web. A number of commenters linked to their personal photos of the scene depicted in the Library of Congress resource. Additional links went to Wikipedia pages and commercial websites. Better visualization of these links may facilitate resource discovery.

Correction and Translations are a correction to a factual error or a translation of one language to another. For example, the place names in one of our sets have changed over time, with the original title now being out of date. The modern place name was supplied by a user, as demonstrated in Figure 3, allowing for search crawlers to index the image under its modern label and presumably make it more likely to be found in a web search. Users occasionally engaged in threaded discussions in the comment stream, correcting or verifying each other’s comments.

The Link In category contains comments that allow for the image to be found via browsing groups of like images. This is another, unofficial, access point to the resource. These groups may be seen as “self-curated” collections. Unlike analog resources, the digital may reside in multiple sets or collections simultaneously.

Specific user-types, such as those searching for images with certain aesthetic may be served by an expanded categorization system. An expanded classification scheme might be produced to include ideas such as: transcription of text in the image, aesthetic judgments, or technical details. Using a classification as outlined above, specific user groups may be given the ability to filter out comments that do not fit their information needs.

Currently, there is no way for users to identify the type of comment they submit to the system. It may be useful to provide users with a user interface that allows them to identify their comments as belonging to one of the categories we propose. These user-supplied classifications could then be used as filters, allowing patrons to view only those comments which are of interest in a particular search. As the comment stream grows ever longer, it might prove beneficial, for example, for the historian to view only historical comments while the artist views only comments categorized as aesthetic.

User-contributed comments and notes are a new phenomenon in the library world. We are looking at a source of rich and valuable content for library resources. Tools and processes should be developed to encourage continued interaction on a deep level between interested and willing patrons and the library.

5 Future Research

A fuller categorization than the four we use in our study of user-contributed metadata is a likely next step. Comparing this new categorization to previous tag-based efforts might prove to expose additional insight into users motivations and practices while annotating images. Differences of annotations for user's own images, versus their annotating of library resources might be useful in showing how ownership and motivation affect metadata creation.

Determining patron and librarian reaction to the use of user-contributed metadata may help guide libraries through the early years of social networking. For example, are librarians and library staff accepting of "crowd-sourced" information? Libraries and museums [9] are experimenting with user-contributions, but there is no widely accepted practice of incorporating this information into catalogs or other data stores. More research is needed to determine the best ways to use this new type of metadata.

Librarians may also be given better tools to act in a gatekeeper role. Notably, the National Archives of the Netherlands instituted a professional review of user contributions [10]. The *steve.museum* initiative has implemented a thumbs-up, thumbs-down voting protocol for moderators in a museum [9]. It may prove beneficial to expand on this sort of voting system. Or, as is found in commenting systems across the Web, patrons may find it useful to self-moderate comments, with some threshold of "thumbs-up" points needed for a particular comment to be considered trusted. However, presumably users are accessing library materials primarily to gather information thus the solicitation of too much additional effort might prove to interfere with users' information seeking behaviors, undermining the very benefits libraries hope to see.

A visualization of the structure of links and connections between resources would be a valuable next step in creating new uses for user-contributions. The user interface may greatly influence how a user both contributes a comment, and how a patron uses it. There are several interesting possibilities in this domain. A timeline view [1], or "places in time" interface may be built to visualize the same location at varying points

in its history. Links to these resources could be detected programmatically and an interface built to allow comparison of two or more images. As the user moves away from the original, or seed, image in the library's holdings they may wind up multiple steps removed from the library. An interface allowing the user to remain within the realm of the library while viewing these external resources might foster additional exploration.

A goal of The Smithsonian's activity with Flickr is to "enhance the documentation and interpretation of our collections using the knowledge, perspectives, and experiences of these audiences" [5]. They also anticipate collecting and storing user contributions, possibly in a catalog record with attribution to the source of the information. While this is an exciting possibility, there are many technical and organizational hurdles to overcome before such a system may be implemented. We reach a point of information overload, where the patron may not have the ability or patience to parse multiple comments, regardless of their value. An analysis of queries and information retrieval techniques used on a set of images may prove to identify the most valuable annotation types [10], allowing a library to solicit comments targeted towards information retrieval.

6 Conclusion

Patrons are no longer passive consumers; they have the desire and ability to enhance library collections for the use of other patrons and professionals. In response to a request for help at the launch of the Flickr Commons, the public has shown they are willing and able to provide detailed and valuable annotations, corrections, and translations for the Library. This community-based effort may provide expertise where a library has none, and increase the number of "catalogers and indexers" identifying and annotating images. Trust and authority of the work is an issue. However, an open exchange of comments, as the discussions have shown, may mitigate the effects of many erroneous postings.

Libraries may be able to leverage "crowd-curated" collections, which take an authoritative resource as a "seed" and build sets and groups without central control. Much like the open-source software movement, we may see open-source collection building, some of these efforts closely affiliated with institutions, others less so. Additional efforts on the development of the user-interface, institutional policies, and privacy/authority assurance are needed before the full use of user-contributed metadata may be realized.

Acknowledgements

We would like to thank the Library of Congress and Flickr teams for initiating the Commons and making their data available via an API.

References

- [1] Allen, R.B.: Timelines as Information System Interfaces. In: Proceedings International Symposium on Digital Libraries, Tsukuba, Japan (August 1995), <http://www.-cis.drexel.edu/faculty/ballen/PAPERS/TL/isdl.pdf>

- [2] Anderson, S., Allen, R.B.: Envisioning the Archival Commons. *The American Archivist* 72(2), 383–400 (Fall/Winter 2009)
- [3] Flickr: Photochrom Travel Views (Set), http://www.flickr.com/photos/library_of_congress/sets/72157612249760312/
- [4] Guy, M., Tonkin, E.: Folksonomies. Tidying Up Tags? *D-Lib Magazine* 12(1) (2006), <http://www.dlib.org/dlib/january06/guy/01guy.html>
- [5] Kalfatovic, M., Kapsalis, E., Spiess, K., Van Camp, A., Edson, M.: Smithsonian team Flickr: A library, archives, and museums collaboration in web 2.0 space. *Archival Science* 8(4), 267–277 (2009)
- [6] Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT 2006, tagging paper, taxonomy, Flickr, academic article, to read. In: *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, HYPERTEXT 2006*, Odense, Denmark, August 22 - 25, pp. 31–40. ACM, New York (2006), doi: <http://doi.acm.org/10.1145/1149941.1149949>
- [7] Marshall, C.C.: No bull, no spin: a comparison of tags with other forms of user metadata. In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2009*, Austin, TX, USA, June 15-19, pp. 241–250. ACM Press, New York (2009), doi: <http://doi.acm.org/10.1145/1555400.1555438>
- [8] Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., Zinkham, H.: For the Common Good: The Library of Congress Flickr Pilot Project. Technical Report. Library of Congress (2008)
- [9] Trant, J.: Tagging, folksonomy and art museums: Early experiments and ongoing research. *Journal of Digital Information* 12(1) (2009)
- [10] van Hooland, S.: Spectator Becomes Annotator: Possibilities Offered by User-Generated Metadata for Image Databases. In: *Proceedings of the CILIP Conference*, University of East Anglia, UK (2006)

An Approach to Content-Based Image Retrieval Based on the Lucene Search Engine Library*

Claudio Gennaro, Giuseppe Amato, Paolo Bolettieri, and Pasquale Savino

ISTI - CNR, Pisa, Italy
{claudio.gennaro,giuseppe.amato,
paolo.bolettieri,pasquale.savino}@isti.cnr.it

Abstract. Content-based image retrieval is becoming a popular way for searching digital libraries as the amount of available multimedia data increases. However, the cost of developing from scratch a robust and reliable system with content-based image retrieval facilities for large databases is quite prohibitive.

In this paper, we propose to exploit an approach to perform approximate similarity search in metric spaces developed by [3,6]. The idea at the basis of these techniques is that when two objects are very close one to each other they 'see' the world around them in the same way. Accordingly, we can use a measure of dissimilarity between the views of the world at different objects, in place of the distance function of the underlying metric space. To employ this idea the low level image features (such as colors and textures) are converted into a textual form and are indexed into the inverted index by means of the Lucene search engine library. The conversion of the features in textual form allows us to employ the Lucene's off-the-shelf indexing and searching abilities with a little implementation effort. In this way, we are able to set up a robust information retrieval system that combines full-text search with content-based image retrieval capabilities.

Keywords: Approximate Similarity Search, Access Methods, Lucene.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval;

1 Introduction

The continuous reduction of the cost of multimedia devices such as cameras, camcorders, and smartphones, is driving the demand for content-based image and video retrieval tools for multimedia digital libraries. Several attempts are currently being made to provide these capabilities, for instance some commercial products like SnapTell (<http://www.snaptell.com>) and Google goggles (<http://www.google.com/mobile/goggles>) have been available for on-line

* This work was partially supported by the VISITO project, funded by the Tuscany region of Italy.

visual search for smartphones. However, the cost of developing and deploying from scratch a robust and reliable system with content-based image retrieval facilities could not be within the range of possibilities for everyone.

In this paper, we would like to approach the problem of similarity search by enhancing the full-text retrieval library Lucene¹ with content-based image retrieval facilities. Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java that is suitable for nearly any application requiring full-text search abilities.

In particular, we use a technique for approximate similarity search when data are represented in generic metric spaces. The metric space approach to similarity search requires the similarity between objects of a database to be measured by means of a distance (dissimilarity) function, which satisfies the metric postulates: positivity, symmetry, identity, and triangle inequality. The advantage of the metric space approach to the data searching is its “extensibility”, allowing us to potentially work for a large number of existing proximity measures as well as many others to be defined in the future. In contrast, many approaches need objects to be represented as vectors and cannot be applied to generic metric spaces.

The basic idea exploited in our approach has been independently introduced by Amato et al [3] and Chavez et al. [6] and consists on observing that two objects x_1 and x_2 are very similar (which in metric spaces means that they are close one to each other), if their view of the surrounding world (their perspective) is similar as well. This implies that, if we take a set of objects from the database and we order them according to their similarity to x_1 and x_2 , the obtained orderings are also similar. Therefore, we can approximatively judge the similarity between any two arbitrary objects x_1 and x_2 , by comparing the ordering, according to their similarity to x_1 and x_2 , of a group of reference objects, instead of using the actual distance function between the two objects.

Clearly, it is possible to find some special examples where very similar (or even identical) orderings correspond to very dissimilar objects. For instance, if reference points are all positioned on a line, two objects that are positioned on another line orthogonal to the first one will produce the same ordering of the reference points, independently of their actual position. However, as it has been proved in [3], even with a random selection of the reference points, the accuracy of this approach is very good.

The structure of the paper is as follows. Section 2 reviews previous work in this field. Section 3 formalizes the idea of searching by using the perspective of the objects. Section 4 shows how this idea can be efficiently supported by the use of the Lucene library. Section 5 proposes a preliminary performance evaluation of the proposed solution.

2 Related Work

Techniques for approximate similarity search can be broadly classified in techniques that exploit space transformations and techniques that reduce the amount

¹ <http://lucene.apache.org>

of data to be accessed. Our approach can be considered a hybrid approach given that, as we will see, even if it is mainly based on a space transformation it also adopts techniques to reduce the amount of data accessed. Extensive literature surveys can be found in [20,3,6].

Among space transformation techniques we mention dimensionality reduction techniques as, for instance, those proposed in [9,14], where the authors propose techniques to approximatively and efficiently compute the inner product between two vectors by exploiting an ad-hoc dimensionality reduction. Space transformation is also used in approximate search techniques based on VA-Files [19] where dimensionality reduction is obtained by quantizing the original data objects. Other techniques that fall in the category of space transformation are FastMap [12], which can be mainly used in vector spaces, and MetricMap [18] suited to metric spaces.

Techniques that reduce the space to be examined basically aim at improving performance by accessing and analyzing less data that is technically needed in order to have a mathematically precise result. These strategies try to infer the most promising portions of the space to be examined or to decide when it might be useless to continue searching for qualifying objects. Many of these techniques were defined exploiting data structures that use an hierarchical decomposition of the space, as for instance the M-Trees [8]. In [15,16] a technique that analyzes the angle formed between objects in a ball region, the center of this region, and a query object, to decide when a region has to be accessed is proposed. This technique can be applied on any data structure based on hierarchical decomposition of the space by means of ball regions when data are represented in a vector space. A technique employing a user-defined parameter as an upper bound on approximation error is presented in [21]. The error parameter is used as an upper bound to the error introduced if a region of the space is not accessed when searching. A technique that retrieves k approximate nearest neighbors of a query object by returning k objects that statistically belong to the set of l ($l \geq k$) actual nearest neighbors of the query object is also presented in [21]. The value l is specified by the user as a fraction of the whole dataset. A technique called *Probably Approximately Correct* (PAC) nearest neighbor search in metric spaces is proposed in [7]. The approach searches the approximate nearest neighbor to a query object guaranteeing that the introduced error is within a user-specified confidence interval. A technique that uses a proximity measure to decide which tree nodes can be pruned, even if their bounding regions overlap the query region, is proposed in [2]. When the proximity of a node's bounding region and the query region is small, the probability that qualifying objects will be found in their intersection is also small. A user-specified parameter is employed as a threshold to decide whether a node should be accessed or not. If the proximity value is below the specified threshold, the node is not promising from a search point of view, and thus not accessed.

One of the early works that suggested the use of inverted index for CBIR is the Viper system [17], in which images are indexed by a huge number of visual features that can either be present or absent in each image, as words in a text.

However, it is not clear if and how this approach is scalable (their experiments are limited on a small collection of images).

In [13] the authors propose a similar approach of integrating the Lucene text search engine and exploiting the principles of our approach to speedup the retrieval. However, to the best of our knowledge, at present time there are no published reports investigating the performance of this solution.

Capitalizing on the work of Amato et al [3], we also use the inverted files in our research. Another similar approach, called MiPai [10], uses a compact prefix tree for estimating the real distance order of the indexed objects with respect to a query. All these above mentioned approaches make use of index methods completely designed and developed from scratch. Although the results of these systems are quite impressive², they probably will not easily move from research prototypes to commercial applications due to the strong effort required to maintain and support such information systems. Consider, for example, Lucene: at the time of this writing, Lucene's core team includes about half a dozen active developers. In addition to the official project developers, Lucene has a fairly large and active technical user community that frequently contributes patches, bug fixes, and new features.

Moreover, only the approach in [10] provides a full-text search on descriptive textual metadata, which is, however, not combined with the content-based similarity search. Our approach instead since it is built on top of Lucene provides complex query processing by combining similarity search with the full-text search.

3 Perspective Based Space Transformation

Let \mathcal{D} be a domain of objects and $d : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ be a metric distance function between objects of \mathcal{D} . Let $RO \subset \mathcal{D}$, be a set of reference objects chosen from \mathcal{D} .

Given an object $x \in \mathcal{D}$, we represent it as the ordering of the reference objects RO according to their distance d from x . More formally, an object $x \in \mathcal{D}$ is represented with $\bar{x} = O(x)$, where $O(x)$ is the vector of ranks of all objects of RO , ordered according to their distance d from x .

We denote the rank in $O(x)$ of a reference object $ro_i \in RO$ as $O_i(x)$. For example, if $O_4(x) = 3$, ro_4 is the 3rd nearest object to x among those in RO . We call \mathcal{D} the domain of the transformed objects. $\forall x \in \mathcal{D}, \bar{x} \in \mathcal{D}$.

Figure 1 exemplifies the transformation process. Figure 1a) sketches a number of reference objects (black points), data objects (white points), and a query object (gray point). Figure 1b) shows the encoding of the data objects in the transformed space. We will use this as a running example throughout the remainder of the paper.

As we anticipated before, we assume that if two objects are very close one to each other, they have a similar view of the space. This means that also the

² <http://mipai.esuli.it/>
<http://mi-file.isti.cnr.it/CophirSearch/>

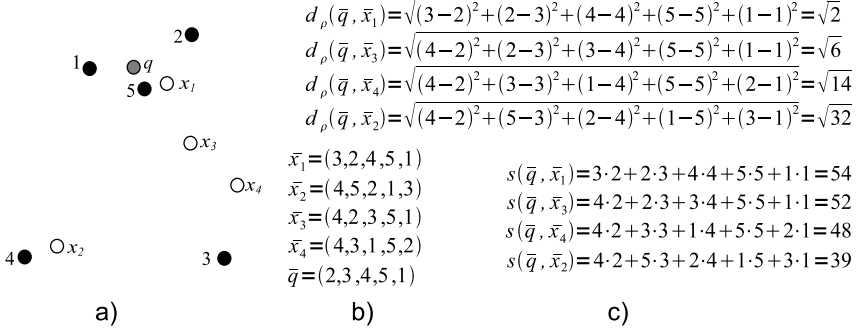


Fig. 1. Example of perspective based space transformation. a) Black points are reference objects; white points are data objects; the gray point is a query. b) Encoding of the data objects in the transformed space. c) Distance d_ρ and similarity s in the transformed space.

orderings of the reference objects according to their distance from the two objects should be similar. There are several standard methods for comparing two ordered lists, such as *Kendall's tau*, the *Spearman Footrule Distance*, and the *Spearman Rho Distance* [11]. In this paper, we concentrate our attention on the latter distance, which is also used in [6]. The reason of this choice (explained later on) is tied to the way standard search engines process the similarity between documents and query. Given two ordered lists $O(x)$ and $O(y)$ ($x, y \in \mathcal{D}$), containing the ranks of all objects of RO , the Spearman Rho Distance d_ρ between $O(x)$ and $O(y)$ is computed as in the following:

$$d_\rho(O(x), O(y)) = \sqrt{\sum_{i=1}^m (O_i(x) - O_i(y))^2}$$

where m is the size of the set of reference objects RO . While this distance measures the degree in which rankings correspond with each other, it is often of interest to measure the degree of noncorrespondence or dissimilarity between two rankings. Spearmans rank correlation s_ρ [11] is probably the best known and most frequently used measure, which is given by:

$$s_\rho(O(x), O(y)) = 1 - \frac{6 \sum (O_i(x) - O_i(y))^2}{m(m^2 - 1)}$$

this measure of dissimilarity, which is normalized in the range $[-1, 1]$, is a monotonic transformation of the above d_ρ distance. Closer s_ρ is to 1, better is the agreement while s_ρ closer to -1 indicates strong agreement in the reverse direction. Since in our case tied ranks do not exist, the Spearmans rank correlation can be assessed by the following product

$$s(O(x), O(y)) = \sum_{i=1}^m O_i(x)O_i(y) \quad (1)$$

This simple scalar product is again a similarity measure that assumes a maximum value if there is a perfect agreement between the two sets of ranks and a minimum value if there is a complete disagreement between the two sets of ranks. With simple mathematical steps, it is possible to prove that s is just a linear transformation of s_ρ . In few words this means that if we use Eq. (1) and Spearman's ρ distance to sort all the objects with respect to an arbitrary query object we obtain the same sequence in inverse order, as Figure 1b) shows.

The transformed domain $\bar{\mathcal{D}}$ and the similarity s can be used to perform approximate similarity search in place of the domain \mathcal{D} and the distance function d . Figure 1c) shows the similarity, computed in the transformed space, of the data objects from the query object. It can easily be seen that it is consistent (it gives the same ordering) with the actual distance of data objects from the query.

Let us suppose that we have a dataset $X \subset \mathcal{D}$ and a query $q \in \mathcal{D}$. Suppose we want to search for the k objects of X nearest to q . An exhaustive approach is to sort the entire dataset X according to the distance from q and to select the first k objects. Let \bar{X} be the dataset in the transformed space and $\bar{q} \in \bar{\mathcal{D}}$ the transformed query. The approximate ordering of X with respect to q can be obtained in $\bar{\mathcal{D}}$ by computing the similarity $s(\bar{q}, \bar{x}), \forall x \in X$. In the following we will show that this ordering can be obtained by representing (indexing) the transformed objects with inverted files and using search algorithms derived from the full-text search area.

In principle, we can index transformed objects with inverted files as follows. Entries (the lexicon) of the inverted file are the objects of RO . The posting list associated with an entry $ro_i \in RO$ is a list of pairs $(x, O_i(x)), x \in X$, that is a list where each object x of the dataset X is associated with the rank of the reference object ro_i in \bar{x} . In other words, each reference object is associated with a list of pairs each referring an object of the dataset and the rank of the reference object in the transformed object's representation. For instance, an entry $(x, 7)$ in the posting list associated with reference object ro_i , indicates that ro_i is the 7th closest object to x among those in RO .

Therefore, the inverted file has the following overall structure:

$$\begin{aligned} ro_1 &\rightarrow ((x_1, O_1(x_1)), \dots, (x_n, O_1(x_n))) \\ &\dots \\ ro_m &\rightarrow ((x_1, O_m(x_1)), \dots, (x_n, O_m(x_n))) \end{aligned}$$

where n is the size of the dataset X and m is the size of the set of reference objects RO .

4 Using Lucene Library

As explained above we would like to exploit the Lucene library as a basic tool for managing the inverted index. The basic idea underlying our approach is to

associate a textual representation to each metric object of the database so that the inverted index produced by Lucene looks like the one presented above and that its built-in similarity function behaves like the Spearman Similarity rank correlation used to compare ordered lists.

Full-text search engines typically use Cosine Similarity to measure the matching degree of the query vector \bar{q} with document vectors \bar{v}_i , i.e.,

$$\cos(\bar{q}, \bar{v}_i) = \frac{\bar{q} \cdot \bar{v}_i}{|\bar{q}| \cdot |\bar{v}_i|},$$

where vectors \bar{q} and \bar{v}_i are usually the term frequency vectors of the documents. However, Lucene does not normalize the product $\bar{q} \cdot \bar{v}_i$ by default, and in practice it uses the simple scalar product as a baseline for similarity evaluation. Therefore the similarity measure computed by Lucene is identical to our similarity s of Eq. (II). However, in our specific case, the vectors will contain the rank of the reference object ro , instead of the term weights of the classical vector space model.

To this purpose, we must generate a text associated to each object to be indexed by Lucene that produces the desiderate posting list. This means that the text to be indexed must use a vocabulary composed of m terms (as the number of reference objects). Thus, we first need to associate a unique textual identifier to each reference object RO . This transformation is denoted by the function $ID(ro_j)$. Then, we define the transformation function $TXT(x_i)$, which is able to associate a textual representation to each object x_i of X . This function first evaluates the vector $O(x_i)$, which is the ordered list containing all objects of RO , ordered according to their distance d from x_i . However, we make a slight modification to our earlier definition of similarity s , we use the complement of the rank value $m + 1 - O_j(x_i)$. For instance if $m = 3$, instead of storing the vector $\bar{x} = (1, 2, 3)$ we will store $x = (3, 2, 1)$. This choice does not affect the value s . Its advantage will be clarified later on. Then for each item $O_j(x_i)$, $TXT(x_i)$ will contain $m + 1 - O_j(x_i)$ repetitions of the string $ID(ro_j)$. Consider the example reported in Figure III, and let us assume $ID(ro_1) = RO1$, $ID(ro_2) = RO2$, etc. The function TXT will generate the following output

```
TXT(x1) = "RO5 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO2 RO1 RO1 RO1 RO3 RO3 RO4"
TXT(x2) = "RO4 RO4 RO4 RO4 RO4 RO3 RO3 RO3 RO3 RO5 RO5 RO5 RO1 RO1 RO2"
TXT(x3) = "RO5 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO2 RO3 RO3 RO3 RO1 RO1 RO4"
TXT(x4) = "RO3 RO3 RO3 RO3 RO3 RO5 RO5 RO5 RO5 RO2 RO2 RO2 RO1 RO1 RO4"
```

and for the query q :

```
TXT(q) = "RO5 RO5 RO5 RO5 RO5 RO1 RO1 RO1 RO1 RO2 RO2 RO2 RO3 RO3 RO4"
```

In order to reduce the size of the inverted index we can exploit the idea of taking just the closest reference objects to represent any object that has to be indexed, exactly as in [3]. Let $k_i \leq m$ be the number of reference objects used for indexing. In this case every object can be represented as $\bar{x} = O(x)$, using k_i nearest reference objects instead of m . Note that, in this case, different objects

will be typically represented by different reference objects, given that different objects will have different neighbor reference objects.

This representation of an object will be clearly smaller than using all reference objects. In addition, this has also the effect of reducing the size of the inverted file. In fact, every object will be just inserted into k_i posting lists, by reducing their size and by also reducing the search cost.

This idea can be extended also to the query, for which we can exploit a number $k_q \leq k_i$ of nearest reference objects. In this case, the advantage is that we can tune the efficiency and the effectiveness of the search at query time.

The benefit of using the complement of the rank becomes now clear if we note that, the entry (x, e) in the posting list, associated to the reference objects ro_i , will take the maximum value $e = k_i$ if it is the first closest object to x , the value $e = k_i - 1$ if it is the second, and so on, down to $e = 0$ in case ro_i does not appear in the first k_i reference objects. In practice, the reference object that has a rank greater than k_i , for an object x , will correspond to the entry $(x, 0)$ in the inverted file. Note that, the value in the posting list 0 usually means that a term is not present in the document and it is treated as a special case in the vector space model. This characteristic allows Lucene to fully exploit its strategies for optimizing the query performance.

To summarize, we have three different parameters to take into account: the number of reference objects m , the number of reference objects used for indexing k_i and the number of reference objects used for querying k_q , with $k_q \leq k_i \leq m$.

5 A Real Application and Performance Evaluation

In this section, we report the results of an experimental evaluation of the proposed method. For both testing and demonstration, we developed a web user interface to perform image content based retrieval on the CoPhIR dataset [5], which consists of 106 millions images, taken from Flickr (www.flickr.com), described by MPEG-7 visual descriptors. Content based retrieval can be performed by using similarity functions of the visual descriptors associated with the images.

We have indexed the whole CoPhIR dataset and for each image, we created five Lucene fields which can be queried separately or in combination. The first field contains the unique identifier of the Flickr image. The second field maintains the textual information taken from title, and tags of the original Flickr image. The other three fields contain the content generated by the TXT function explained above for searching on three different pre-combined visual features. In particular, in order to support content based search, the CoPhIR project extracted several MPEG-7 visual descriptors from each image, three descriptors for describing the colors (SCD, CSD, and CLD) and two for describing textures (EHD and HTD). We have indexed three different aggregations of those descriptors, the first one combining the three color descriptors, the second one combining the two texture descriptors, and the third one combining all five descriptors. In this way we leave the possibility to the user to search for colors and textures independently or to search all the descriptors together. The weights used for aggregating the descriptors are the ones suggested in [4].

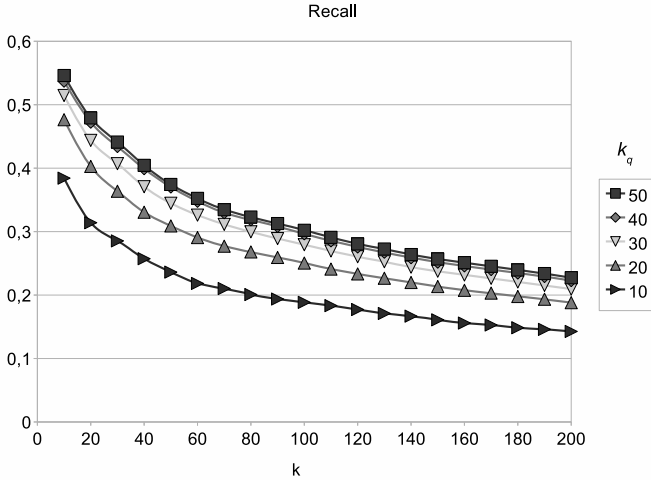


Fig. 2. Recall varying the number k for different values of k_q parameter

From one of the authors' home page³ it is possible to find a link to the demo web application of the developed search engine. From that page it is possible to perform a full-text search, a similarity search starting from one of the random selected images. Besides the three types of visual similarities, thanks to the search functionality of Lucene, it also provides complex query processing by combining any of the three types of similarity search with the full-text search on descriptive metadata.

We conducted our experiments using the combination of all visual descriptors, with 20,000 reference objects and by setting $k_i = 50$ during the indexing. We used the measure of the recall to assess the accuracy of the method. Specifically, given a query object q , the recall is defined as $R = \frac{\#(S \cap S^A)}{\#S}$, where S and S^A are the ordering of the k closest objects to q found respectively by the exact similarity and by the proposed method. In practice, we compare the efficacy of our solution with an algorithm that exploits a sequential scan of the whole database. The comparison was made at the same conditions, using only the similarity obtained as combination of all five MPEG-7 descriptors, without exploiting the textual content. For this purpose 100 queries were randomly selected from the database. Results are shown in Figure 2. The graphs show the recall varying the number of items retrieved k for various options of the $k_q \leq k$. The performance are very similar to the one presented in [10], where the same dataset was used.

Figure 3 shows the average query processing time as function of k_q . As expected, the search cost increases with the size of k_q , and become unacceptable for $k_q > 20$. This performance can be easily improved if the architecture consists of multiple Lucene indexes, since Lucene search framework includes parallel and multi-threads search facilities. Our index consists of ten separated Lucene

³ <http://www.nmis.isti.cnr.it/gennaro/>

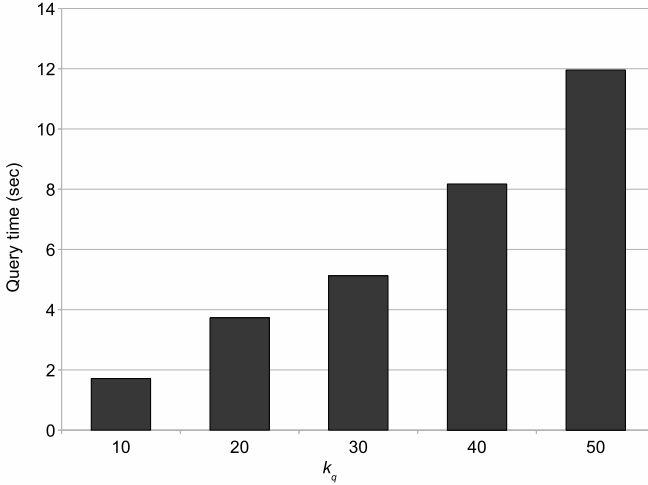


Fig. 3. Query time for different values of k_q parameter

indexes each one including about 1/10 of the whole dataset. If the indexes reside on different physical disks, we may obtain performance improvements; however, in our tests conducted with a single physical disk, the performance with multi-thread search was slightly better than with a single-thread search. The total space occupation of the Lucene indexes is 530GB, which means about 5.24KB for each image record.

6 Conclusions and Future Work

In this paper we presented an approach to approximate similarity search in metric spaces based on a space transformation that relies on the idea of perspective from a data point. We proved through a concrete implementation that the proposed approach has clear advantages over other methods existing in literature in terms of easiness in implementation. A major characteristic of the proposed technique is that it can be implemented by using inverted files, thus capitalizing on existing software investments. There are still some issues that are worth of investigations to further improve this technique. In order to reduce the search cost it is possible to increase the number of reference objects. This solution has however the disadvantage to expand the indexing time, which took about one week for the 106 millions images. To deal with this problem, we could make use of an existing metric access method for efficient similarity search for indexing the reference objects in main memory. In this way, during database indexing phase, we should speed up the search of the nearest reference objects to the object to be inserted, by invoking the k nearest neighbor search provided by the metric access method.

References

1. Spearman's rank correlation coefficient, http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient
2. Amato, G., Rabitti, F., Savino, P., Zezula, P.: Region proximity in metric spaces and its use for approximate similarity search. *ACM Trans. Inf. Syst.* 21(2), 192–227 (2003)
3. Amato, G., Savino, P.: Approximate similarity search in metric spaces using inverted files. In: *Proceedings of the 3rd International Conference on Scalable Information Systems (InfoScale 2008)*, pp. 1–10. ICST (2008)
4. Batko, M., Kohoutkova, P., Novak, D.: Cophir image collection under the microscope. In: *International Workshop on Similarity Search and Applications*, pp. 47–54 (2009)
5. Bolettieri, P., Esuli, A., Falchi, F., Lucchese, C., Perego, R., Rabitti, F.: Enabling content-based image retrieval in very large digital libraries. In: *Second Workshop on Very Large Digital Libraries (VLDB 2009)*, DELOS, pp. 43–50 (2009)
6. Chavez, E., Figueroa, K., Navarro, G.: Effective proximity retrieval by ordering permutations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1647–1658 (2007)
7. Ciaccia, P., Patella, M.: Pac nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces. In: *ICDE*, pp. 244–255 (2000)
8. Ciaccia, P., Patella, M., Zezula, P.: M-tree: An efficient access method for similarity search in metric spaces. In: Jarke, M., Carey, M.J., Dittrich, K.R., Lochovsky, F.H., Loucopoulos, P., Jeusfeld, M.A. (eds.) *VLDB 1997, Proceedings of 23rd International Conference on Very Large Data Bases*, Athens, Greece, August 25–29, pp. 426–435. Morgan Kaufmann, San Francisco (1997)
9. Egecioglu, Ö., Ferhatosmanoglu, H.: Dimensionality reduction and similarity computation by inner product approximations. In: *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2000)*, McLean, Virginia, USA, November 6–11, pp. 219–226. ACM Press, New York (2000)
10. Esuli, A.: Pp-index: Using permutation prefixes for efficient and scalable approximate similarity search. In: *Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval (LSDS-IR 2009)*, pp. 17–24 (2009)
11. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top-k lists. *SIAM J. of Discrete Math.* 17(1), 134–160 (2003)
12. Faloutsos, C., Lin, K.-I.: FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In: Carey, M.J., Schneider, D.A. (eds.) *Proceedings of the 18th ACM International Conference on Management of Data (SIGMOD 1995)*, San Jose, California, USA, May 22–25, pp. 163–174. ACM Press, New York (1995)
13. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: *MM 2008: Proceeding of the 16th ACM International Conference on Multimedia*, pp. 1085–1088. ACM, New York (2008)
14. Ogras, Ü.Y., Ferhatosmanoglu, H.: Dimensionality reduction using magnitude and shape approximations. In: *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM 2003)*, New Orleans, Louisiana, USA, November 3–8, pp. 99–107. ACM Press, New York (2003)
15. Pramanik, S., Alexander, S., Li, J.: An efficient searching algorithm for approximate nearest neighbor queries in high dimensions. In: *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999)*,

- Florence, Italy, June 7-11, vol. 1. IEEE Computer Society Press, Los Alamitos (1999)
16. Pramanik, S., Li, J., Ruan, J., Bhattacharjee, S.K.: Efficient search scheme for very large image databases. In: Beretta, G.B., Schettini, R. (eds.) Proceedings of the International Society for Optical Engineering (SPIE) on Internet Imaging, San Jose, California, USA, January 26, vol. 3964, pp. 79–90. The International Society for Optical Engineering (December 1999)
 17. Squire, D.M., Müller, W., Müller, H., Pun, T.: Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters* 21(13-14), 1193–1198 (2000); Selected Papers from The 11th Scandinavian Conference on Image
 18. Wang, X., Wang, J.T.-L., Lin, K.-I., Shasha, D., Shapiro, B.A., Zhang, K.: An index structure for data mining and clustering. In: *Knowledge and Information Systems*, vol. 2, pp. 161–184. Springer, Heidelberg (2000)
 19. Weber, R., Böhm, K.: Trading quality for time with nearest neighbor search. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) *EDBT 2000. LNCS*, vol. 1777, p. 21. Springer, Heidelberg (2000)
 20. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach. In: *Advances in Database Systems*, vol. 32. Springer, Heidelberg (2006)
 21. Zezula, P., Savino, P., Amato, G., Rabitti, F.: Approximate similarity retrieval with m-trees. *VLDB J* 7(4), 275–293 (1998)

Appendix: Technical Details

From the implementation point of view, in order to produce a correct inverted file, we must instruct Lucene not to employ the tf-idf-based weighting scheme. For this reason the `Similarity` class of Lucene was overridden to return raw term frequencies, as in the following

```
public class NoIDFSimilarity extends Similarity {
    public float idf(int docFreq, int numDocs) {
        return 1f;
    }
    public float tf(float freq) {
        return freq;
    }
}
```

However, the new class `NoIDFSimilarity` must be used only during search, when Lucene applies the tf-idf weighting scheme. Therefore, let `s` be our instance of the Lucene `IndexSearcher` class, we have to pass an instance of `NoIDFSimilarity` by the call `s.setSimilarity(new NoIDFSimilarity())`.

Evaluation Constructs for Visual Video Summaries

Stina Westman

Department of Media Technology
Aalto University School of Science and Technology
P.O. Box 15500 00076 Aalto Finland
stina.westman@tkk.fi

Abstract. This paper reports on a user-centered evaluation of visual video summaries. We evaluated four types of summaries (fastforward, user-controlled fastforward, scene clips and storyboard) with a set of existing performance and satisfaction measures. We further conducted a repertory grid elicitation with our participants gathering evaluation constructs related to both video summary content and controls. Results showed a lack of correlation between performance and satisfaction measures. User-supplied evaluation constructs were shown to span both the performance and satisfaction dimensions of the video summary evaluation space. Most constructs achieved moderate to good inter-rater agreement in a consequent survey.

Keywords: video summarization, evaluation measures, repertory grid.

1 Introduction

The goal of video summarization is to create video surrogates, i.e. summaries that facilitate access to video content. Video summaries are kinds of metadata akin to abstracts that stand for the full video object and are useful for the purposes of browsing and making sense of retrieved video objects. Summaries may be used to find out if a specific video is the one we want to watch or to browse through a collection of videos. Visual video summaries may be employed in result sets where people aim to make relevance judgments about whether to look further or download the video. They could also assist in the retrieval process by functioning as navigation aids through a collection or a video stream.

Evaluating the quality of video summaries in these different contexts presents significant challenges. Within the interactive evaluation paradigm, summaries may be evaluated in a task-specific manner (e.g. does the summary help in making relevance assessments) which aims for ecological validity [1]. Another viewpoint into the evaluation of video summaries focuses on their effectiveness in retaining the gist of the original video (i.e. their informative function) [2]. Useful measures for human performance related to video summaries are being investigated [3]. The full range of features that viewers use to evaluate video summaries is unknown and evaluation methodologies are still under development.

We conducted a user study on video summarization combining existing performance and satisfaction measures with an exploratory analysis in the form of repertory

grid analysis. We aimed to see how currently suggested performance and satisfaction evaluation measures of video summaries relate and what additional evaluation constructs for visual video summaries could be elicited from users. This study contributes to the understanding of the relative advantages and disadvantages of different types of video summaries, and describes a user-centered methodological approach for collecting users' criteria for summary evaluation via the repertory grid.

2 Related Work

2.1 Types of Video Summaries

Video summaries are created in order to give viewers access to video content without having to watch the entire video. This enables users to browse large video collections and otherwise interact with video sequences in a non-linear manner [4]. For most applications, video summaries serve two functions: an indicative function, where the summary is used to indicate what topics are contained in the original video; and an informative function, where summaries are used to cover the information in the full video as much as possible, subject to summary length [2].

Different types of video and user data are used to create video summaries. Truong and Venkatesh [4] present a review and classification of techniques for video abstraction, i.e. the mechanisms for creating video summaries. Various types of summaries exist, e.g. textual keywords, static keyframe mosaics and dynamic video skims. Different modalities contribute differently to the information users gain from video summaries. Marchionini et al [5] found that while audio summaries were thought to be less ambiguous and provide keywords, visual summaries provided the overall gist of the video topic. Video summaries may be shown to users only once, combined with unlimited pausing [6] or both the capability to pause and replay [5]. Users have expressed wishes towards more control over the summary playback [7].

2.2 Methods for Evaluating Video Summaries

There are several approaches to evaluating video summaries. Intrinsic evaluations are direct evaluations of the summaries while extrinsic evaluations investigate how the summaries help in some predetermined task [6]. Both types of evaluations have been conducted yet no standard methodology has emerged. Evaluations may be based on participants watching the full video and choosing the best option out of alternative summaries [8]. Participants may be asked to directly judge the quality of summaries or asked questions which are used to calculate summary performance [9]. It is also possible to conduct evaluations relative to an optimal summary [10].

Wildemuth et al [7,11] presented a framework for the evaluation of video summaries with four classes of variables thought to influence performance and satisfaction in video summarization: user tasks, user characteristics, video characteristics, and summary characteristics. The tasks or human performance measures were expanded upon by Marchionini [3] who defined two classes of cognitive measures based on perceptual and conceptual facets of video viewing. Recognition measures evaluate subject's recall about what they saw. Object recognition may be evaluated by participants' ability to recognize by textual or visual stimuli (keyframes) objects seen in the

summary. Action recognition is evaluated similarly by visual stimuli (video clips). Inference measures evaluate how subjects understood the aboutness of what they saw, i.e. the gist of the video. Linguistic gist is evaluated by the accuracy and coverage of a written summary of the video or by having participants select the best summary for it. For visual gist evaluation participants are to select objects that “belong” in the video represented by the summary from still images not seen in the summary but present in the original.

Measures of user satisfaction pertaining to video summaries are still largely lacking [7]. Li et al [12] employed measures related to visual and audio quality, semantic continuity, ability to browse content, how well the summary summarized the content, and the degree to which the summary replaced the need to see the original. Ma et al [9] had participants evaluate summaries according to their enjoyability (if perceptually enjoyable video segments were selected) and informativeness (capability of maintaining content coverage while reducing redundancy). In the TRECVID rushes evaluation campaign subjective measures were the fraction of important segments from the full video included, how easy it was to find the desired content, and how much redundancy the summary contained [6]. Recently a set of subjective measures on usability, usefulness, enjoyment, and engagement have been used in addition to eliciting free-form written comments from participants [5].

2.3 Repertory Grid Analysis

The repertory grid is a cognitive mapping technique designed to reveal the personal constructs individuals use to structure and interpret phenomena [13]. Unlike a conventional questionnaire, the repertory grid utilizes constructs that originate from the participant. In the field of information science the repertory grid technique has been used to gain insight into e.g. information assets [14], document types [15] and search engines [16]. The repertory grid technique contains three major components: elements, constructs and links [13]. Elements represent aspects considered important within the domain and may be supplied by the researcher or elicited from the participants. Constructs represent participants’ interpretations of the elements, i.e. labels that participants use to make sense of the elements. These are commonly elicited from participants in the form of bipolar labels (e.g. easy to use - hard to use). Various methods may be employed to link elements and constructs. Most often rating scales are used to differentiate between elements on each elicited construct. Repertory grids may be analyzed as individual grids or across several grids via multivariate methods (e.g. cluster analysis, factor analysis, correspondence analysis).

In our study, the elements are different video summary types and the constructs related to these are elicited from one set of participants. Another set of participants then proceeds to rate the elements according to the constructs.

3 Methods and Analysis

We conducted a user test on video summarization to answer the following questions: How do current performance and satisfaction measures for video summaries relate

to each other? What additional constructs for summary evaluation can be elicited from viewers?

Our focus was on informative visual summaries. Research results indicate that when only one modality with automatic summary generation is used, visual summaries fare better than audio [5]. We evaluate the ability of the summaries to convey information about the original video. This does not preclude their indicative function which we assess in a subjective manner.

3.1 Participants and Material

We recruited 28 participants (12 female) through postings in university newsgroups. They were engineering students between ages 20 and 37 (mean 24.8 years). None had any previous experience with video summaries.

We selected four documentary videos of similar topics from the Open Video (OV) Project for the study: 1) A New Horizon, segment 5¹ (water), 2) Challenge at Glen Canyon, segment 5² (dam), 3) Exotic Terrane, segment 10³ (volcano), 4) Hurricane Force - A Costal Perspective, segment 2⁴ (hurricane).

Four types of summaries (Table 1) were produced for each video, about 6% of the length of the originals. We devised both static and moving summaries, and allowed for user controls in one summary type. The fastforward rate is relatively slow due to the focus on informative summaries. We allow two viewings per summary (or respective time) due to findings that users in the unlimited viewing condition view visual summaries multiple times [5].

Table 1. Summary types and their details

Type	Content	Controls	Time
Storyboard	Keyframes from OV (if necessary supplemented with manually selected frames from other shots), to total 16 keyframes	All keyframes visible at once	16 s
Scene clips	Compilation of 500 ms clips of original video starting from storyboard keyframes	Shown twice	16 s
Fastforward [7]	Every 16 th frame of the original video resampled	Shown twice	16 s
User-controlled fastforward	Every 16 th frame of the original video. Based on suggestions for mechanisms for user control of display speed in fastforwards [7]	A drag and drop functionality for faster playback and pausing	16 s + 2 s to access controls

3.2 Measures

The visual recognition and inference measures suggested in [3,17] were used as performance measures. We further added an inference measure related to action based on the importance of actions and activities in visual gisting [18]. *Action inference* was

¹ <http://www.open-video.org/details.php?videoid=689, 1:59>

² <http://www.open-video.org/details.php?videoid=566, 2:02>

³ <http://www.open-video.org/details.php?videoid=728, 2:13>

⁴ <http://www.open-video.org/details.php?videoid=837, 2:13>

evaluated by the ability to infer actions in the video based on the summary. *Object recognition* and *object inference* (visual gist by visual stimulus [3]) were evaluated by having participants indicate for 18 frames whether they:

1. saw the frame in the summary (recognition) [6 correct frames]
2. did not see but thought it belonged to the original video (inference) [6 frames]
3. did not see and thought it did not belong to the original video [6 frames; 3 from other segments of the same video, 3 from unrelated videos]

Recognition and inference questions were thus integrated into a multiple choice screen. *Action recognition* and action inference were evaluated in an analogous manner from a set of six 2 s clips. Inference was also evaluated by gathering free text descriptions of summaries and full videos. Results on these are presented elsewhere.

At the end of the summary trials participants re-answered the recognition questions for the last video. This enabled us to gather a (between-subjects) baseline for the full videos. There was no statistically significant difference between the videos so absolute scores have been used. As the fastforward and user-controlled fastforward summaries included nearly all shots, object and action inference were disregarded for these and only recognition is reported.

Satisfaction questions (5-point Likert scale) were derived and combined from literature:

1. The summary was easy to understand [6]
2. The summary was enjoyable [9]
3. The summary was informative [9]
4. The summary was interesting
5. The summary was coherent [19]
6. The summary represented the video well [2]
7. The summary would aid in deciding whether to watch the full video [2]
8. The summary would replace watching the full video [19]
9. The summary would be useful in browsing video archives [1]

3.3 Procedure

An online system based on PHP/MySQL and JavaScript was used to administer the study and collect data through a counterbalancing of summary type and video instance order (a within-subjects Greco-Latin design of 4 blocks). The following procedure was used, with phases 2-7 repeated for the four summary type trials:

1. Introduction to test and practice viewing all summary types and questions
2. Watch video summary
3. Describe video based on summary
4. Answer performance questions and satisfaction questions 1-5
5. Watch corresponding full video
6. Describe full video
7. Answer satisfaction questions 6-9 comparing summary and full video
8. Exit interview and repertory grid elicitation or survey

3.4 Elicitation of Further Evaluation Constructs

We gathered further constructs for the evaluation of visual video summaries by two methods. First, in the exit interview participants were encouraged to reflect across summary types and video instances and answering e.g. Which summary type did you prefer? Which did you think was most informative? What did you pay attention to when evaluating the summary? Was some video content more interesting than others?

Second, we used the repertory grid technique to study the users' mental model of the summaries and to elicit subjectively valid evaluation constructs. Two pilot participants and first eight test participants took part in a repertory grid interview designed to elicit constructs. At the end of the test they were presented with a screen displaying all four summaries they had seen in trials. They were to name features shared between any two of those summaries or features distinguishing them. The interviewer noted down the constructs (e.g. usability) and asked further questions to establish its endpoints on the semantic differential scale (easy to use vs. difficult to use). The items pooled from the ten interviews revealed 20 distinct constructs. A single participant contributed between 1 and 5 constructs. No new constructs emerged after the 8th participant. The 20 items were listed in a survey instrument, designed to gather rating data for all summary types on a 5-point scale. The rest of the participants ($n=20$) filled out the repertory grid survey with these 20 constructs as items. We used this survey data to evaluate the summary types and the constructs gathered.

4 Results

Results on summary performance and satisfaction are presented first, followed by elicited constructs. Then, relations between different types of measures are explored.

4.1 Performance and Satisfaction by Summary Type

Summary type had an effect on object recognition ($F=9.448$, $df=3$, $p<.001$) (Figure 1). According to Games-Howell post hoc tests object recognition was better for scene clips than for either type of fastforwards ($p<.001$). The enjoyability of the summary types differed ($F=5.050$, $df=3$, $p<.01$) as storyboards were more enjoyable than

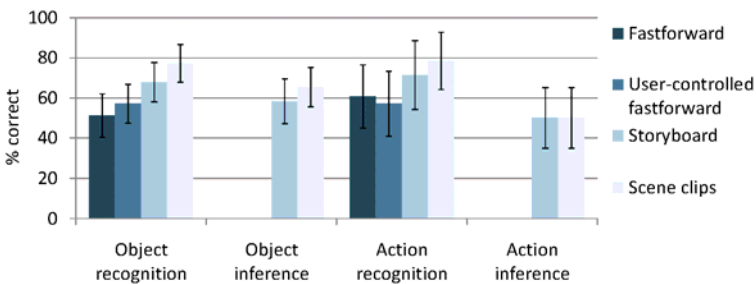


Fig. 1. Average percentage of correct answers to recognition and inference questions. Error bars in all figures denote one standard deviation from mean.

fastforwards ($p < .01$) (Figure 2). Differences were also found in how well the summary represented ($F = 3.005$, $df = 3$, $p = .03$) or could replace the full video ($F = 5.782$, $df = 3$, $p < .001$) (Figure 3). Scene clips were more representative than storyboards ($p = .03$). Scene clips ($p < .01$) and user-controllable fastforwards ($p < .01$) would be better able to replace full video than storyboards. Video instance had effects on the ease of understanding ($F = 12.802$, $df = 3$, $p < .001$), enjoyability ($F = 4.862$, $p < .01$) and informativeness ($F = 5.220$, $p < .01$) of the summaries. We found no effects of gender, age (under 24 vs. older), or presentation order.

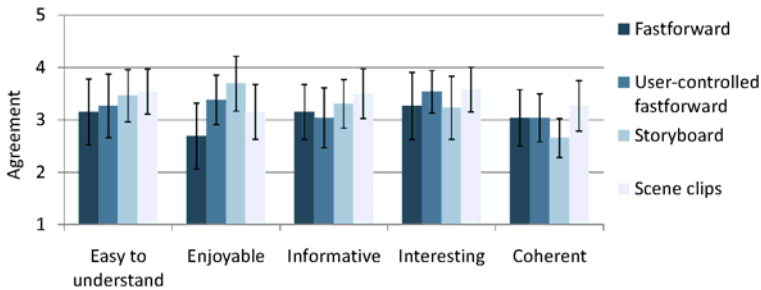


Fig. 2. Average agreement (1=completely disagree, 5=completely agree) to satisfaction questions

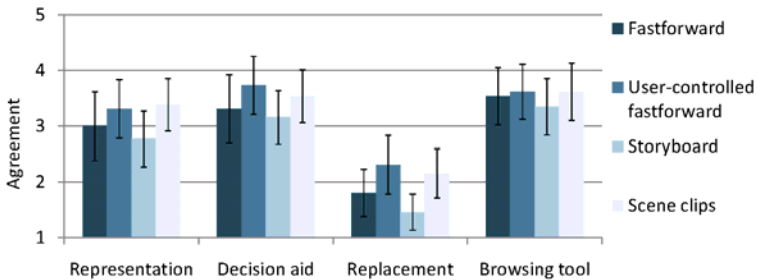


Fig. 3. Average agreement to additional satisfaction questions

4.2 Summary Qualities and Constructs

The fastforwards and scene clips were preferred as summary types and thought to be most informative by most participants (Table 2). Most preferred the hurricane ($n = 12$) or volcano (9) video with fewer mentioning dam (5) or water (2) as their favorite. The ease of understanding a summary ($F = 9.845$, $df = 1$, $p < .01$) and its interestingness ($F = 11.627$, $df = 1$, $p < .001$) were higher for preferred content. The evaluation constructs elicited from the first set of participants are given in Table 3. To evaluate the inter-rater reliability of the constructs we calculated an average linearly-weighted Cohen's kappa [20] for each construct across all rater pairs in the second set of participants. A kappa value of 0 denotes no agreement and value of 1 perfect agreement.

Table 2. Summary type evaluation in post-test survey (P=preferred, I=most informative)

Type	P (n)	I (n)	Reasons cited
Fast-forward	8	10	Best conveyance of plot, best overall picture, displays most information in a short time, shows whole video, offers continuity
User-controlled ff	5	7	One can choose what to focus on, ability to go back to interesting spots, most control over the summary playback, most useful, nothing gets missed
Storyboard	3	2	Could see overall picture at once, got an idea about different parts of video, could watch what one wanted, ability to spend more time on a frame
Scene clips	12	9	Clearest overall idea of content, best usability, easy to follow, closest to original, best for recognizing scenes, time to process

Table 3. Constructs and their average linearly-weighted Cohen's kappa (k)

Construct	Scale endpoints 1-5	k
Adjustable.speed	Standard speed - Adjustable speed	.86
Automated	Can stop playback - Cannot stop playback	.80
Controllable	Automated playback - Controllable playback	.59
Multiple.images	Shows one frame at a time - Multiple frames at a time	.58
From.video	Constructed from still images - Constructed from video	.53
Small.space	Takes up lot of screen space - Takes up little screen space	.52
Moving.imagery	Still images - Moving images	.49
Chosen.shots	Covers whole content of the video - Contains selected spots	.48
Slow.playback	Presentation speed high - Presentation speed slow	.47
Focusable	Have to focus on what is shown - Can select focus	.42
Increased.speed	Normal paced video - Increased pace video	.41
Reviewable	Difficult to view a part closer - Easy to view a part closer	.37
Fidelity	Shots missing - Repeats the content of the video with fidelity	.36
Continuity	One has to piece up the continuation - Shows the continuation	.36
Usability	Difficult to use - Easy to use	.30
Skipping	Have to watch completely - Can skip parts	.26
Searchable	Difficult to locate a certain spot - Easy to locate a certain spot	.19
True.content	Content deformed - Content not deformed	.19
Overall.picture	Does not provide an overview - Provides an overview of video	.08
Fast.aboutness	Slow to find out what video is about - Fast to find out what video is about	.04

We conducted a correspondence analysis on the construct rating data to show interrelations between 1) summary types, 2) constructs and 3) summary types and constructs (Figure 4). Dimension one ("coverage", related to summary content) was significantly correlated with e.g. ability to choose what to focus on and inclusion of chosen parts from the video. Dimension two ("effort", related to summary control) was correlated with e.g. adjustability of playback speed and low usability. These dimensions accounted for 60% of variance in individuals' summary rating data.

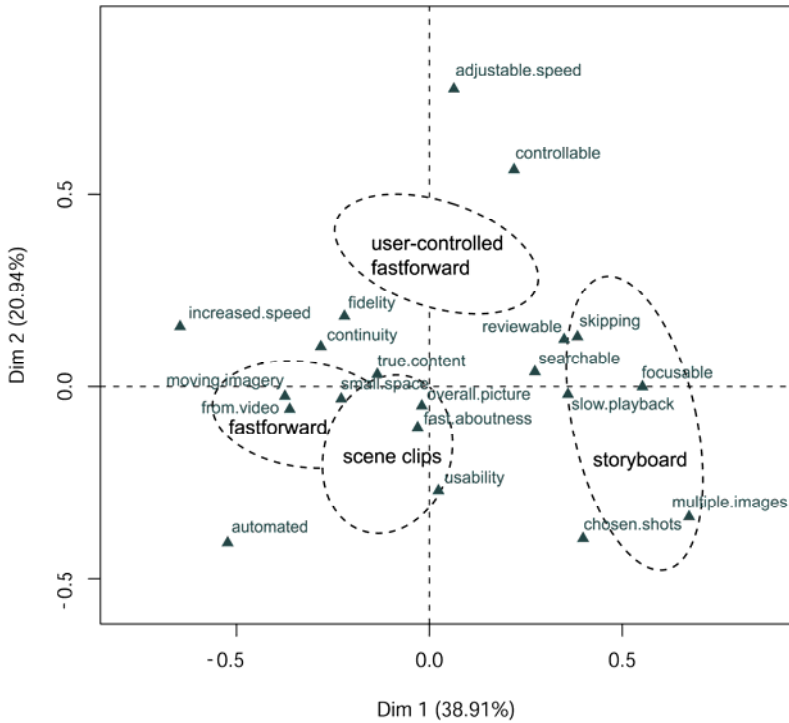


Fig. 4. Correspondence analysis results. Confidence ellipses encircle 95% of rating points for individual summaries.

4.3 Connections between Measures

In a correlation circle (Figure 5) the coordinates of each measure represent its correlations with the axes. While the explained variance in the two-dimensional plot is low (34%), the visualization serves to highlight connections between the types of measures (performance, satisfaction, constructs). Dimension one builds upon the satisfaction measures while performance measures correlate more with dimension two. Some constructs (e.g. focusable, searchable) correlate with dimension one and others (e.g. fidelity, continuity) correlate more with dimension two.

We found strong correlations between the performance measures: the two recognition measures correlated ($r=.415$, $n=112$, $p<.001$) as did the inference measures ($r=.378$, $n=56$, $p<.001$) as. There was correlation between object-oriented measures ($r=.238$, $n=56$, $p<.05$) but almost none between action-oriented measures. All satisfaction measures correlated strongly ($p<.01$), and only the first set (from Figure 2) has been visualized here. There were weaker correlations between performance and satisfaction measures. Object recognition correlated with informativeness ($r=.136$, $n=112$, $p=.076$). Action inference correlated with coherence ($r=.239$, $n=56$, $p=.038$) and ease of understanding ($r=.202$, $n=56$, $p=.067$).

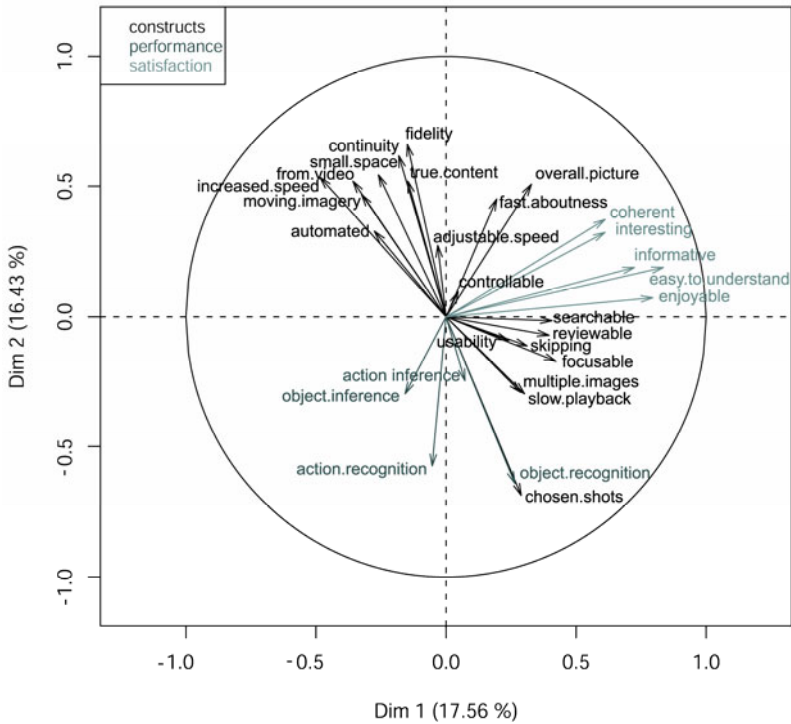


Fig. 5. Correlations between different measures visualized in a correlation circle

5 Discussion

Current measures showed little difference between still images and respective video clips (storyboard vs. scene clips) or between automated and user-controlled summaries (fastforwards vs. user-controlled fastforwards). The performance of the summary types differed only for object recognition. Satisfaction evaluations differed on enjoyability, representativeness and ability to replace original. Storyboards were on par with dynamic summaries on performance and were most enjoyable but lacked representativeness and ability to replace the original. Most participants preferred moving summaries, stating that storyboards had too low fidelity to be considered good video summaries. Storyboards were thus informative but not indicative.

We tested a measure of action inference which correlated with subjective evaluations of summary quality as measured by coherence and informativeness. Keyframes and clips could be combined as test stimuli for visual inference measurement in contexts where both object and action inference are important.

The evaluation constructs elicited show that users are able to distinguish between multiple qualities of summaries. The correspondence analysis of the construct data shows that both content (modality, selection of shots) and presentation (user controls)

of metadata surrogates [21] are recognized and reflected upon by video summary users. The lack of major correlations between performance and satisfaction measures means both are needed in evaluation setups. User-supplied constructs mapped between the performance and satisfaction dimensions in the correlation plot indicating that user-supplied constructs reflect both evaluation modes. Similar issues have been raised in free comments in previous studies but were now gathered as reliable rating scale items. The low degree of variance explained by the two-dimensional correlation plot reflects the multidimensional issue of quality of visual video summaries. Multivariate methods have been used early on in video summary evaluation [22] and have the potential to highlight important dimensions in video summaries.

Constructs related to summaries' ability to convey the aboutness or overall picture of the video polarized participants resulting in low kappa values. For some, the constant frame rate of the fastforwards represented continuity. Others preferred the normal-paced scene clips whose continuous shots were also useful for object recognition performance. Results show that displaying the context of shots makes video summaries more useful in the retrieval process [23]. For our users the issue of gaining an overall, coherent picture of the video contents seemed central.

Due to the relatively high frame rate used we omitted the data on inference measures for fastforwards and user-controlled fastforwards. We had, in order to keep the test procedure was identical for all summary types, displayed the inference answer option of "did not see but is included in the original video" for all trials. The inclusion of the non-relevant answer alternative might have created a negative bias towards fastforward types of summaries. Users did select more recognition choices for the fastforwards and user-controllable fastforwards ($\chi^2=9.93$, $df=3$, $p=.02$) so they did not select the inference alternatives in a forced manner despite this issue.

Christel [1] warns against the generalization of results as some types of summaries match certain genres better than others. In this study satisfaction was affected by the video instance. Differences were related to ease of understanding, enjoyability and informativeness of the summaries. We also found an effect of content preference on interestingness. This was not paralleled by any effect of video instance but arose from participants' individual preferences. The constructs elicited here might, to some degree, be specific to these video and summary types, thus needing further evaluation.

6 Conclusions and Future Work

We have presented results on user-centered evaluation constructs for visual video summaries. It remains important to develop and utilize both performance and satisfaction measures as these do not correlate in a straightforward manner. Correlation within measure groups (performance, satisfaction) enables the streamlining of evaluation procedures. The constructs obtained here through repertory grid analysis spanned both performance and satisfaction dimensions and showed acceptable inter-assessor agreement. They could be used as basis when developing novel evaluation measures

for visual video summaries. We are interested in evaluating these user-supplied constructs as measures in task-focused studies on video summaries, e.g. in situations of browsing and relevance assessments.

References

1. Christel, M.C.: Evaluation and user studies with respect to video summarization and browsing. In: *Multimedia Content Analysis, Management and Retrieval*, pp. 196–210 (2006)
2. Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D., Delp, E.J.: Automated video program summarization using speech transcripts. *IEEE T. Multimedia* 8, 775–791 (2006)
3. Marchionini, G.: Human performance measures for video retrieval. In: *8th ACM International Workshop on Multimedia Information Retrieval*, pp. 307–312 (2006)
4. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 3 (2007)
5. Marchionini, G., Song, Y., Farrell, R.: Multimedia surrogates for video gisting: Towards combining spoken words and imagery. *Inform. Process. Manag.* 45, 615–630 (2009)
6. Over, P., Smeaton, A.F., Kelly, P.: The trecvid 2007 BBC rushes summarization evaluation pilot. In: *TRECVID Workshop on Video Summarization*, pp. 1–15 (2007)
7. Wildemuth, B.M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., Gruss, R.: How Fast Is Too Fast? Evaluating Fast Forward Surrogates for Digital Video. In: *3rd ACM/IEEE-CS JCDL*, pp. 221–230 (2003)
8. Corchs, S., Ciocca, G., Schettini, R.: Video summarization using a neurodynamical model of visual attention. In: *IEEE 6th Workshop on Multimedia Signal Processing*, pp. 71–74 (2004)
9. Ma, Y.-F., Lu, L., Zhang, H.-J., Li, M.: A user attention model for video summarization. In: *ACM Multimedia, ACM MM* (2002)
10. Guironnet, M., Pellerin, D., Guyader, N., Ladret, P.: Video Summarization Based on Camera Motion and a Subjective Evaluation Method. *J. Image Video Processing* (2007)
11. Marchionini, G., Wildemuth, B.M., Geisler, G.: The Open Video Digital Library: A Möbius strip of research and practice. *J. Am. Soc. Inf. Sci. Technol.* 57, 1629–1643 (2006)
12. Li, Y., Narayanan, S., Kuo, C.-C.J.: Movie Content Analysis, Indexing and Skimming via Multimodal Information. In: Rosenfeld, A., Doermann, D., Dementhon, D. (eds.) *Video Mining*. Kluwer Academic Publishers, Dordrecht (2003)
13. Tan, F.B., Hunter, M.G.: The repertory grid technique: A method for the study of cognition in information systems. *MIS Quarterly* 26, 39–57 (2002)
14. Oppenheim, C., Stenson, J., Wilson, R.M.S.: Studies on Information as an Asset II: Repertory Grid. *J. Inform. Sci.* 29, 419–432 (2003)
15. Dillon, A., McKnight, C.: Towards a classification of text types: a repertory grid approach. *Int. J. Man-Mach. Stud.* 33, 623–636 (1990)
16. Crudge, S.E., Johnson, F.C.: Using the repertory grid and laddering technique to determine the user's evaluative model of search engines. *J. Doc.* 63, 259–280 (2004)
17. Yang, M., Wildemuth, B.M., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., Webster, C.: Measures of User Performance in Video Retrieval Research. UNC (SILS) Technical Report TR-2003-02 (2003)

18. Yang, M., Marchionini, G.: Deciphering visual gist and its implications for video retrieval and interface design. In: ACM CHI, pp. 1877–1880 (2005)
19. Sundaram, H., Chang, S.-F.: Condensing computable scenes using visual complexity and film syntax analysis. In: IEEE ICME (2001)
20. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 213–220 (1968)
21. Balatsoukas, P., Morris, A., O'Brien, A.: An evaluation framework of user interaction with metadata surrogates. *J. Inform. Sci.* 35, 321–339 (2009)
22. Goodrum, A.: Multidimensional scaling of video surrogates. *J. Am. Soc. Inf. Sci. Technol.* 52, 174–182 (2001)
23. Wildemuth, B.M., Russell, T., Ward, T.J., Marchionini, G., Oh, S.: The Influence of Context and Interactivity on Video Browsing. UNC (SILS) Technical Report TR 2006-01 (2006)

Visual Expression for Organizing and Accessing Music Collections in MusicWiz

Konstantinos Meintanis and Frank M. Shipman

Dept. of Computer Science and Engineering, Texas A&M University
College Station, TX 77843-3112
{kam2959, shipman}@cse.tamu.edu

Abstract. Music services, media players and managers provide support for content classification and access based on filtering metadata values, statistics of access, and user ratings. This approach fails to capture characteristics of mood and personal history that are often the deciding factor when creating personal playlists and collections in music. This paper presents MusicWiz, a music management environment that combines traditional metadata with spatial hypertext-based expression and automatically extracted characteristics of music to generate personalized associations between songs. MusicWiz's similarity inference engine combines the personal expression in the workspace with assessments of similarity based on the artists, other metadata, lyrics, and the audio signal to make suggestions and to generate playlists. An evaluation of MusicWiz with and without the workspace and suggestion capabilities showed significant differences for organizing and playlist creation tasks. The workspace features were more valuable for organizing tasks while the suggestion features had more value for playlist creation activities.

Keywords: Spatial hypertext, media managers, music recommendation.

1 Introduction

The majority of people manage their personal music collections via explicit attributes. Metadata values like the artist, the composer and the genre are used extensively in presenting the collection. An organization based on these classification schemes can be understood by most users as it is based on well-defined criteria. The common metadata fields attached to music are valuable for providing context-free information about the music – the artist of a recording does not change between playbacks – but are not necessarily the music characteristics that express what users really seek. How can users pick music that reminds them of high school, or of college, or of particular family members or friends? While they can add metadata fields, they rarely do so because the potential value is outweighed by the overhead of expression [21], especially when their expression involves interpretation that is likely to change over time, such as the feelings and memories triggered by listening to the music. Retrieving songs based on ratings and access statistics can reliably detect music of preference but not necessarily music desired for a specific setting.

Our belief when starting this work was that, while managing and using personal music collections benefits from the consistency and accuracy of explicit expression, it would also benefit by the less restrictive information resulting from implicit forms of expression. This combination is embodied in MusicWiz, an environment that supports associating songs based on personal feelings and memories.

MusicWiz consists of several components that collect, process and display information about music: a preprocessor that prepares the songs for the analysis phase, a database for storing the music attributes and their similarity values, a playback module, an organization interface and a similarity inference engine. This paper focuses on the last two components; particularly, its support of free-form, non-verbal expression, its approach for assessing music relatedness and the presentation of songs similar to those being considered by the user. The next section provides a high-level description of our approach. Section 3 overviews related work. A formative study exploring the use of non-verbal expression for music characteristics is discussed in Section 4. Section 5 presents the MusicWiz interface and similarity inference engine, followed by a comparative usage study and conclusions (Sections 6 & 7).

2 Approach

Limitations of metadata in describing imprecise and complicated concepts make it difficult for current media applications to provide access based on an individual's feelings or memories. Our approach is to combine explicit and implicit information about songs along with a medium that facilitates the expression of personal interpretation. By combining these three forms of information, we aim to provide access that is closer to how individuals perceive, remember or feel music.

Explicit and implicit information is used by many systems. Explicit information consists of the metadata attached to the music and any explicit ratings and labels provided by users. Implicit information is anything not explicitly assigned to a song that can be derived by analyzing its content (e.g. audio signal and lyrics), access patterns and statistics, and its membership in a collection or category. Implicit information is not constrained by the limitations of formal representation. It can contain details about how a song is harmonically and dynamically structured (e.g. music identification based on frequency components and volume evolution over time), personal preferences (e.g. what are the attributes/style of the songs the user listens to frequently or puts in the same playlist), and collection management practices (e.g. what are the attributes of songs placed in the same collection and what distinguishes songs in different collections).

Personal interpretations have the potential to add to the above information because the feelings and memories that songs trigger are often vague or tacit. Nor can they be easily deduced from community logs – the songs that remind us of our first love are unlikely to be identified by mining the playlists of others. Non-verbal media, such as expressions that involve color-coding items and/or placing them in piles or lists, reduce the need for the user to come up with descriptors for ill-defined and evolving concepts and categories. While ambiguous, visual express provides a representation of the user's feelings. This project explores the inclusion of such expression.

3 Related Work

Related work in music management falls into three main categories: systems that use explicit and implicit information to provide access to music, visualizations providing access to music libraries, and music digital library infrastructure.

In the first category we have media players like the Windows Media Player, the QuickTime Player, and the Real Player as well as media managers like the iTunes, the Media Monkey, the Media Catalog Studio, and the Songs-DB. In most of them, the interaction with the music collection is based on filtering metadata, statistics of use, access patterns and ratings associated with the songs. In addition to ratings, the Genius application in iTunes uses collaborative filtering (CF) for recommendation of similar music. Recommendations are based on the user's personal music library, the songs in the iTunes Store and the "anonymously-gathered knowledge from millions of other iTunes users" (<http://www.apple.com/sg/pr/library/2008/09/09itunes.html>). Instead of inferring similarity, Pandora internet radio employs a large number of music experts to classify and associate songs according to a pool of 400 musical attributes (<http://www.pandora.com/corporate/>). Given a song, an artist or keyword, the system searches for overlaps in the human assigned attributes and returns a playlist with the best matches. Instead of using human experts to identify relevant music, Last.fm uses CF to generate recommendations. Last.fm users have a detailed taste profile that is constantly updated according to their music selections and feedback (<http://www.last.fm/help/faq?category=99>). Users can also recommend artists, songs or albums directly to others (individuals or groups) and can listen to "recommendation radio" featuring the music that has been recommended to them.

Visualizations for accessing music can be classified into those where the relatedness between the songs is static and cannot be updated and those that allow users to dynamically change the similarity assessments based on their interactions. Most music-collection visualization tools fall into the first category. Examples include Islands of Music [17], Globe of Music [11], and nepTune [9]. In contrast, MusicSim [2] provides a "graph view" to display songs as 2-D objects clustered according to their content-based similarity and previous user feedback. Songs are positioned relative to the cluster center according to their similarity to the centroid and other neighboring songs. Their location is not fixed and users can reposition them (within the same cluster or to another one) according to their own perception of similarity and hence influencing system's assessments of related music. In Musicream [6], songs are automatically grouped and color-coded based on the similarity of their mood. The songs stream down, one after the other, from taps on the top of the screen. Users can select falling songs for listening or use the "similarity-based sticking function" to "stick" music they want to listen to into a playlist.

Music digital libraries have explored architectures and alternative access mechanisms for browsing and searching collections. In the VocalSearch music search engine [18], users can query the database using text-based lyrics and by singing the melody or music notation. Hanna and colleagues [7] propose a retrieval system that is based on the similarity in the chord progressions. Chord progression comparison is also used by Kuo and Shan [10] in combination with instrument, volume and highest pitch information for music classification based on the melody style. Tsai and Wang [22] propose a music digital library architecture where songs are classified and

accessed based on vocal-related information and more specifically the voice characteristics of their singers. Recognizing the importance of associating user-generated opinions to music objects, Downie and Hu [4] analyze online music reviews to find the kind of terms people use to comment negatively or positively. Bischoff and colleagues [1] have developed algorithms for creating mood (opinion) and theme (occasion) classifiers as well as genre predictors based on user annotations (tags extracted from Last.fm) and lyrics. The Son of Blinker (SOB) system and its underlying Networked Environment for Music Analysis (NEMA) [5] provide a visualization of (machine-generated) audio-based classifications (e.g. genre, mood, artist, etc.) synchronized to the music to indicate the evolution of moods or genres.

4 Preliminary Study

We performed a formative study to identify potential opportunities and issues concerning the use of visual workspaces for organizing music. In this study, twelve participants (10 male, 2 female, age 24-38) were given 60 minutes to organize 100 songs in a spatial hypertext environment and then were asked to create three playlists for different events within 30 minutes. Participants were encouraged to “think out of the box” of traditional metadata and to express their own interpretation of the music.

Spatial hypertext environments are designed to reduce the overhead of user expression for ambiguous and difficult to describe concepts. A spatial hypertext consists of a set of information objects with visual attributes (e.g., color, border width) and spatial layout (e.g., lists, piles) to indicate relations between information entities [12]. Study participants used the Visual Knowledge Builder 2 (VKB2) [20], a general-purpose spatial hypertext system where information is placed into a hierarchy of two-dimensional visual workspaces called collections. By providing a wide range of modifiable visual attributes and the ability to organize materials in space, users can express a variety of relations and their strengths without having to verbally express the meaning and degree of relations. Figure 1 shows one of the resulting music organizations. The color and border width variations are the result of the user’s expression. VKB displays the title and artist for each song. When the cursor lingers over the border of an object, additional metadata is shown in a popup. VKB plays the first 10 seconds of an audio file while the mouse cursor lingers over an audio object. Double-clicking on an object would play the music file in Windows Media Player.

A pre-task questionnaire found that all participants had previous experience in organizing songs. Only one participant was satisfied with the playlist creation techniques based on metadata and usage statistics found in most commercial and freeware software. 83% (10 of 12) create their playlists manually through browsing and dragging-and-dropping songs into their players.

Because the music was pre-selected, participants were confronted with both songs they knew and songs they did not know. This affected the resulting organizations, as seen in Figure 1. This participant divided the songs into those he knew and those he did not. The unknown songs were organized based on the participant’s opinion about the artist (“generally like the artist”, “neutral about the artist”). The songs he knew were grouped based on personal assessments of the music (“like but hard to listen to”,

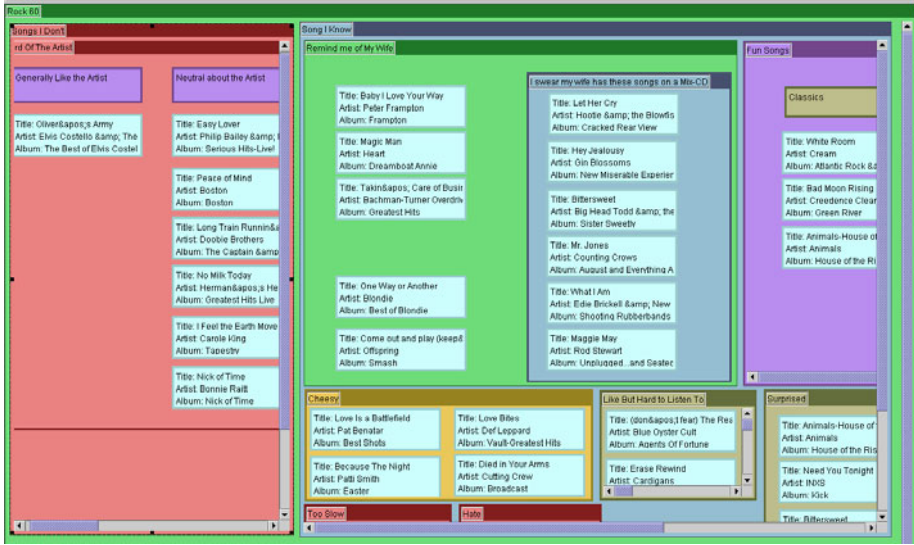


Fig. 1. Organization using categories, subcategories and labels

“cheesy”, “hate”, “fun songs”, and “too slow”) and associations the music had for the participant (“remind me of my wife”). Some of these categories had further subcategories such as the “I swear my wife has these songs on a mix-CD” under “remind me of my wife” and “classics” under “fun songs”.

The categories used to arrange music were examined to determine if the visual workspace was being used for expression found in existing systems. Eight participants used positive and/or negative descriptors of their preference for songs. Musical characteristics were common with seven participants including descriptors of musical features including the mood (“serious”, “peaceful”, “calm”, “aggressive”) and genre (“funky”, “ethnic/folk”, “punk”). A point of interest is that the mood and genre descriptions of several participants were finer-grained than found when represented as metadata. Three organizations included labels to the contexts in which they would want to listen to music (“programming”, “gloomy day”, “off to sleep”). The personal interpretation of these spaces in this controlled study is consistent with the “idiosyncratic genres” found in Cunningham’s ethnographic study [3].

With regard to creating playlists, participants indicated creating playlists requires the selection of items that sound good and fit well together. 83% of the participants reported that music dynamics were important and some participants (25%, 3 of 12) said that they formed playlists based on the lyrics (e.g. what the lyrics say).

While participants liked using visual expression for organizing music, they still wanted to interact with collections based on the metadata values and the explicit associations between songs. Participants expressed an appreciation of the visibility of metadata in the music objects as it supported an initial assessment of what could possibly sound good together. Participants also liked the music preview feature for helping users identify songs they already knew. However, playing the first 10 seconds was not sufficient for assessing new songs. For more details of this study see [15].

5 MusicWiz Design

Results from the preliminary study indicated the potential value and limitations of non-verbal expression in music management, informing the design of MusicWiz. MusicWiz's architecture consists of two major components: an interface for interacting with a music collection and a similarity inference engine for assessing music relatedness based on a combination of explicit and implicit information about music and the user's personal interpretation in the interface.

5.1 MusicWiz Interface

MusicWiz's interface includes an information workspace similar to that of VKB2 alongside a traditional file-system view of the collection, a region for MusicWiz to present search results and suggestions, a pane for playlist creation, and controls for music playback (see Figure 2). As in VKB2, songs in the MusicWiz workspace are represented as rectangular objects that are placed in a hierarchy of collections. Users can modify their background and border color, the border thickness, the font and style of the text, and the size of the component. Participants in the preliminary study appreciated the direct access to metadata values as it enabled a first gross classification without having to listen to the music. Music objects in the workspace display the ID3 tag values, the file location and the song's lyrics. Users can also create plain text objects as annotations/labels in the workspace and resize the views and panels to match their collection and needs.

In the preliminary study, participants complained about not being able to play songs from within VKB. MusicWiz provides full playback functionality directly and extended the preview ability of VKB to include 22-second long multi-phrase music summaries as previews. A study comparing the multi-phrase summaries to the introductions of songs indicated a strong preference for multi-phrase summaries [14].

One of the complaints participants in the preliminary study had was the unavailability of metadata-based and location-based hierarchical views of the music collection. Such conventional classifications are more consistent and resistant to change over time which means that they can be used as a safe starting and reference point for building less conventional organizations. MusicWiz provides a file-system view of the music in the collection alongside the spatial hypertext workspace.

Beneath the file-system view is the pane presenting search results and songs that are similar to the currently selected songs in the file-system view and / or the workspace. Users can also search their music collection based on a wide range of attributes including metadata values, lyrics (occurrence of a specific phrase or set of phrases), and content based attributes (e.g. the beat, brightness, or key of the songs).

A separate tab provides users with easy access to songs similar to the current selection, helping users select music for playlists without having to perform many searches. The similar songs are presented in a tree providing songs that are indirectly similar to the current selection. Users can specify how many similar songs are shown for each branch as well as the depth of the tree. User also select and weight the similarity metrics used by the inference engine, as described in the next section. The creation and population of a playlist in MusicWiz can be done either manually or automatically. Creating a playlist manually is a process of dragging songs from other

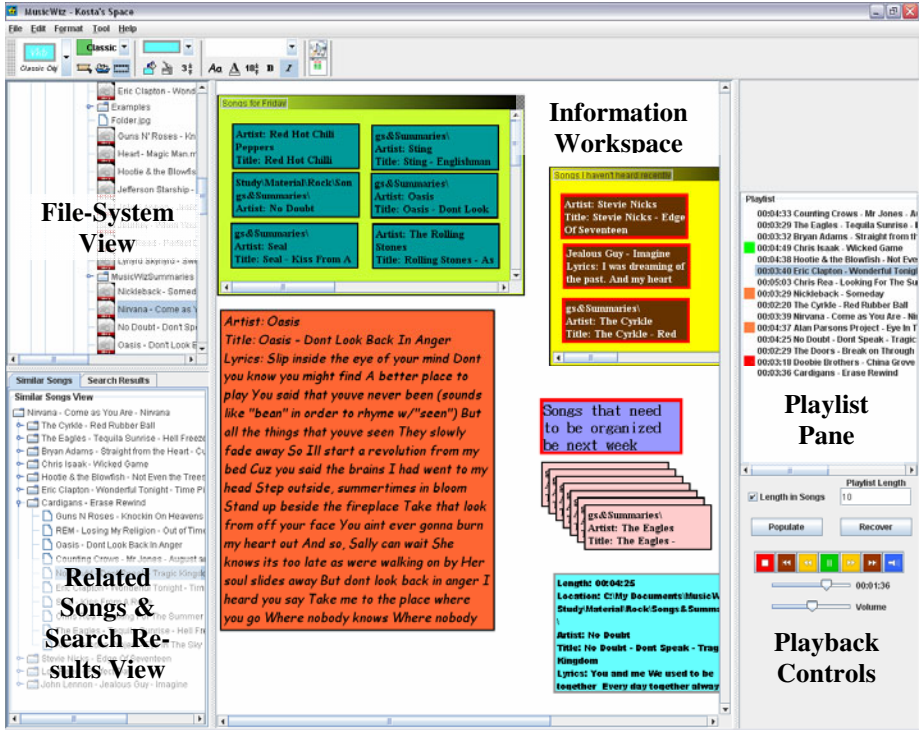


Fig. 2. MusicWiz’s interface combines a tree file-system view, a workspace, and an area for search results and related music as well as a playlist pane and integrated playback controls

panes of the interface into the playlist. The system provides two basic modes for system-assisted playlist creation: filter-based and similarity-based. In filter-based mode, MusicWiz selects music based on the ID3 tags. In similarity-based mode, MusicWiz selects music that is similar to the songs in the current playlist. Users can control the types of data included in the similarity assessment.

5.2 MusicWiz Similarity Inference Engine

MusicWiz’s similarity inference engine supports access to the music collection through relatedness. Music can be related in many ways. It can have similar melody or sound features, be by the same artist, and have lyrics that share common themes, convey a similar mood or feeling, or be viewed as similar due to personal history.

The inference engine consists of several modules that are responsible for extracting, representing, and comparing information about the songs to assess their relatedness. Currently, modules for processing and comparing artists, metadata, audio signals, lyrics, and workspace expression are included. Each of these modules produces an assessment of relatedness (a normalized value ranging from 0 to 1) that is integrated by the inference engine to assess the overall similarity of songs.

Artist Module. The artist module computes similarity based on the results of an analysis of the co-occurrence of 400 popular artists in playlists from the OpenNap

file-sharing network and the Art of the Mix website (<http://www.artofthemix.org/index.asp>) made available by Dan Ellis at Columbia. If the similarity for the artists is unavailable, the metadata module evaluates the proximity of the artist names.

Metadata Module. The metadata module compares the title, artist, genre, album-name, and year of the songs as well as the file-system path where they are stored. It uses a distance metric that combines the Soundex [8] and the Monge-Elkan [16] algorithms to identify transliterated or misspelled names for individual words and variations in the word order. Once individual metadata fields have been compared, these assessments are averaged for an overall rating of the similarity of two songs.

Audio Signal Module. The audio signal module relies on digital signal processing (DSP) to compare the beat (tempo), brightness (centroid), and pitch (fundamental frequency) of the two audio signals. The greater the distance in these features, the less likely two songs are perceived as being of similar style or mood. The beat similarity is the calculated to be the lower number of beats per minute divided by the higher. The brightness of a song is strongly related to the centroid of the sound. Centroid is a popular psycho-acoustical feature that quantifies the mean frequency range of the signal in relation to the amplitude. In simple terms, it measures the position in Hz of the center of mass of the signal's frequency spectrum. The higher the centroid is, the brighter the signal sounds to the human ears. The module averages the similarity of the songs' average brightness and their maximum brightness to generate the overall brightness similarity. The pitch is a subjective psychophysical attribute of the sound that has to do with how humans perceive musical tones. It is strongly related to the harmonics of a sound and especially the lowest of them known as fundamental frequency. To calculate the pitch similarity, MusicWiz determines the similarity in the five most frequent fundamental frequencies of the songs, calculates the similarity in the potential key that the songs are written, and compares their starting note. After the values for beat, brightness and pitch similarity have been calculated, MusicWiz determines the overall audio signal similarity to be the average of these three.

Lyrics Module. The lyrics module uses textual analysis of the lyrics to identify similar songs. Lyrics are scraped from a pool of popular websites and stored in the local database for either display in the objects of the workspace or processing and comparison. To assess the lyrical similarity of two songs, MusicWiz generates their term vectors and calculates their cosine similarity [19].

Workspace Expression Module. The workspace expression module employs the same spatial parser used in VKB2 (and earlier in the systems VIKI and VKB) [13] to identify relations between the components of the information workspace based on their visual attributes and spatial layout. The output of the MusicWiz parser is a forest of trees. Each tree represents a recognized spatial structure in the workspace. Song objects part of a structure are leafs in the tree and can be at different levels in a tree. The workspace expression module defines the similarity of two songs based on the length of the path between the songs in the tree, if they are in the same tree.

Overall Similarity. The overall similarity between songs is computed based on the assessments of the above modules. Users can select which modules are active, which

audio features are included, and weight each module's contribution. The default weights are set to provide a reasonable assessment of overall similarity.

6 Evaluation

There are two central hypotheses in the design of MusicWiz: that a freeform workspace and that suggestions based on a multi-faceted similarity metric will be valuable for collection management and use. A comparative study examining the effects of these features included twenty volunteers. The participants were students, faculty, and staff from a variety of domains. The participants were mainly males (16 of 20) under 36 years old (18 of 20).

6.1 Study Task

Participants were given a collection of fifty classic rock songs and asked to complete three tasks: one requiring classification of the music and two involving searching and similarity assessment. In the first task, songs had to be organized into sub-collections according to participants' own categorization scheme. There was no restriction in the number, the type or the content of the sub-collections that had to be created. In the second task, participants had to form three twenty-minute long playlists based on three different moods or occasions of their choice using songs from their sub-collections. In the third task, participants had to form three six-song long playlists using their sub-collections, but this time the content of each playlist had to be similar (or related) to a specific song (not from the fifty of the original collection). Participants had unlimited time to complete all tasks.

To assess the contribution of MusicWiz's workspace and similarity suggestions, participants were divided (equally and randomly) into four groups of system use. The first group (no workspace / no suggestions) had to complete the three tasks using MusicWiz's browsing, search, and playback functionality and using Windows Explorer folders to form the sub-collections and playlists. Participants in the second group (no workspace / with suggestions) used the same features as the participants of the first group but also received suggestions from the similarity inference engine. In the third group (with workspace / no suggestions), participants had to perform the tasks using the features available in the first group but used the MusicWiz workspace to create the collections and playlists. Finally, the participants of the last group (with workspace / with suggestions) had all MusicWiz features.

The use of Windows Explorer folders for the "no workspace" conditions rather than a music management application (like iTunes) was based on a combination of evidence that many people use the file system to manage their collections and that it would be the most familiar interface across participants. The demographic data found that only 25% of the participants in the preliminary study and 30% of the participants in the current study used specialized software for organizing their music collection.

6.2 Study Results

Results from the study include quantitative data about participant activity (e.g. time taken for tasks), participant assessments of tasks and support via 7-point Likert-scale responses (i.e. "strongly disagree" to "strongly agree"), and open ended comments.

6.2.1 Task One: Classification of Music

The average time taken to organize the music collection varied across the different configurations with the average completion time of task one for Group 1 being 46.2 minutes (longest of the groups, standard deviation $s = 11.0$) while the respective time for Group 3 was just 28 minutes (shortest of the groups, $s = 13.0$). This difference approaches statistical significance ($\alpha = 0.1$, p -value = 0.06) according to the Wilcoxon test (Anova could not be used as the data set does not fulfill the conditions of normality). The average time for Group 2 of 44 minutes ($s = 15.5$), was close to Group 1 and the completion time of 31 minutes ($s = 16.0$) for Group 4 participants was similar to that for Group 3, indicating that the workspace made the organization task more efficient while the suggestions neither helped nor hindered organization.

These time results are supported by participants' assessments on the quality of support they were provided by the system. For the statement "I had enough support to effortlessly/quickly organize the songs the way I wanted", the average rating for Group 1 was 4.4 ($s = 1.5$), the lowest among the four groups. Group 3 rated the support they had as 5.6 ($s = 0.9$). Group 4 was the most satisfied with an average of 6.2 ($s = 0.8$) while Group 2 answered with a 5.4 ($s = 1.9$). One factor could be that Group 3 and Group 4 had quicker access to music via the song previews available in the workspace. This interpretation is supported by the comments of several participants about the value of song previews for quickly assessing music.

The most unexpected result came from the two workspace groups diverging in their rating for the statement "it will be easy for someone else to understand the way I organized the songs". Group 4 was the most positive (5.8 avg., $s = 1.1$). Surprisingly, Group 3 (4.2 avg., $s = 1.6$) was lowest. Group 1 and Group 2 agreed on a 5.4 avg. ($s = 0.5$ and 1.5 respectively). There appears to be an interaction between the workspace features and the suggestion features. One interpretation is that Group 3, using the MusicWiz workspace without suggestions, created organizations that made sense to them but they lacked confidence they would make sense to anyone else. Group 4 may be deriving confirmation and support for the organization from the suggestions.

6.2.2 Task Two and Three Playlist Creation

The time to complete the playlist creation tasks showed no significant differences across the conditions while Likert responses were fairly similar for playlist creation tasks as they were for the organization task. When rating the statement "I had enough support to effortlessly/quickly browse and select the songs" for their playlists, the participants with suggestions (Group 2 and Group 4) were the most satisfied, rating it 6.2 and over, followed closely by the participants in Group 3 (5.8 and 5.6 avg., $s = 0.8$ and 1.9 in tasks 2 and 3 respectively). The participants in Group 1 were barely positive when evaluating system support with 4.8 and 4.4 average ($s = 1.6$ and 1.5) on the two tasks. Table 1 is a summary of the averages for all tasks and groups – the most positive assessment for each statement/task is shown in bold.

When asked about the statement "I had enough support to browse and find the songs I was interested in", the Group 1 was again the least positive (4.8 and 4.6 avg., $s = 1.5$ for both tasks two and three). Group 4 strongly agreed on the sufficiency of their system (6.8 and 6.4 avg., $s = 0.4$ and 0.9) with Group 2 almost as positive (6 and 6.4 avg., $s = 0.7$ and 0.5). Without suggestions but with the workspace, Group 3 rated the support they had as a 5.4 and 6 ($s = 1.1$ and 1.4) in the two playlist-creation tasks.

Table 1. Avg. of 7-point Likert ratings for playlist creation – higher values are more positive

Statement	Task	Group1	Group2	Group3	Group4
support for quick selection	Two	4.8	6.2	5.8	6.2
	Three	4.4	6.8	5.6	6.2
support for finding	Two	4.8	6	5.4	6.8
	Three	4.6	6.4	6	6.4
enjoyed doing task	Two	5.2	6	5.8	6.4
	Three	5.2	5.8	6.4	6.6

When asked about their enjoyment during creating playlists, Group 4 responded most positively (6.4 and 6.6 avg., $s = 0.9$ and 0.5). Again, Group 1 was the most negative. Overall, these results imply that suggestions are more important for supporting playlist creation than the workspace, although the workspace enhanced participants' satisfaction and enjoyment as well as their perceptions of support.

7 Discussion and Conclusions

We have explored the potential for visual and spatial expression in the context of organizing and selecting from a music collection. The results of the preliminary study showed that there are benefits and weaknesses in organizing personal music collections based on the context-independent metadata found in current tools and the malleable personalized interpretation found in spatial hypertext. Participants found visual expression facilitated their interpretation of mood, memories, and musical dynamics. Yet, participants also indicated that the lack of views of their collection based on traditional metadata made it more difficult to locate songs that they knew they wanted. The study also found that users view metadata as insufficient for expressing their desires for playlists.

The MusicWiz personal music management environment was designed based on this feedback. It combines the easily expressed interpretations of music found in spatial hypertext workspaces with the predictable and consistent explicit descriptions found in current metadata-based applications. In MusicWiz, users can associate songs by manipulating their representation in the workspace, can browse and retrieve music based on its lyrics, metadata values and melody features, and can navigate the collection according to the similarity of its content. MusicWiz's similarity inference engine combines independent analyses of similarity based on metadata, audio signal, lyrics, and the user's workspace organization to provide suggestions.

A study comparing MusicWiz with and without suggestions and the workspace showed the workspace was most valuable for organizational tasks while suggestions were more valuable for playlist creation tasks. Regardless of the type of the task they had to perform, the participants using the MusicWiz system with both suggestions and the workspace were the most positive about their overall experience. Further studies are needed to determine whether interactions between the two features indicate user reliance on suggestions or that suggestions are being used to confirm user beliefs. Overall, the study confirmed the value of a workspace for efficient personal expression combined with easy access to similarity-based song selection.

References

1. Bischoff, K., Firan, C., Nejdil, W., Paiu, R.: How do you feel about “dancing queen”? : deriving mood & theme annotations from user tags. In: Proc. of JCDL, pp. 285–294 (2009)
2. Chen, Y., Butz, A.: Musicsim: integrating audio analysis and user feedback in an interactive music browsing UI. In: Proc. of IUI, pp. 429–434 (2009)
3. Cunningham, J.S., Jones, M., Jones, S.: Organizing digital music for use: an examination of personal music collections. In: Proc. of ISMIR, pp. 447–454 (2004)
4. Downie, S., Hu, X.: Review mining for music digital libraries: phase II. In: Proc. of JCDL, pp. 196–197 (2006)
5. Downie, S., West, K., Hu, X.: Dynamic classification explorer for music digital libraries. In: Proc. of JCDL, pp. 422 (2008)
6. Goto, M., Goto, T.: Musicream: New Music Playback Interface for Streaming, Sticking, Sorting, and Recalling Musical Pieces. In: Proc. of ISMIR, pp. 404–411 (2005)
7. Hanna, P., Robine, M., Rocher, T.: An alignment based system for chord sequence retrieval. In: Proc. of JCDL, pp. 101–104 (2009)
8. Knuth, D.: *The Art of Computer Programming. Sorting and Searching*, vol. 3. Addison-Wesley, Reading (1973)
9. Knees, P., Schedl, M., Pohle, T., Widmer, G.: Exploring Music Collections in Virtual Landscapes. *IEEE Multimedia* 14(3), 46–54 (2007)
10. Kuo, F., Shan, M.: Looking for new, not known music only: music retrieval by melody style. In: Proc of JCDL, pp. 243–251 (2004)
11. Leitich, S., Topf, M.: Globe of Music - Music library visualization using GeoSOM. In: Proc. of ISMIR (2007)
12. Marshall, C., Shipman, F.: Spatial Hypertext: Designing for Change. *Comm. of the ACM* 38(8), 88–97 (1995)
13. Marshall, C., Shipman, F., Coombs, J.: VIKI: Spatial Hypertext Supporting Emergent Structure. In: Proc. of ACM ECHT, pp. 13–23 (1994)
14. Meintanis, K.A., Shipman, F.: Creating and Evaluating Multi-Phrase Music Summaries. In: Proc. of ISMIR (2008)
15. Meintanis, K., Shipman, F.: Expressing Personal Interpretations of Music Collections in Spatial Hypertext. *Journal of Digital Information* 10(3) (2009), <http://journals.tdl.org/jodi>
16. Monge, E.A., Elkan, C.: The field matching problem: Algorithms and applications. In: Proc. of the Second Conference on Knowledge Discovery and Data Mining, pp. 267–270 (1996)
17. Pampalk, E., Rauber, A., Merkl, D.: Content-based organization and visualization of music archives. In: Proc. of Multimedia, pp. 570–579 (2002)
18. Pardo, B., Little, D., Jiang, R., Livni, H., Han, J.: The vocalsearch music search engine. In: Proc. of JCDL, pp. 430 (2008)
19. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. *Comm. of the ACM* 18(11), 613–620
20. Shipman, F., Hsieh, H., Airhart, R., Maloor, P., Moore, J.M.: Visual Knowledge Builder: Second Generation Spatial Hypertext. In: Proc. of Hypertext, pp. 113–122 (2001)
21. Shipman, F., Marshall, C.C.: Formality Considered Harmful. *Computer-Supported Cooperative Work* 8(4), 333–352 (1999)
22. Tsai, W., Wang, H.: On the extraction of vocal-related information to facilitate the management of popular music collections. In: Proc. of JCDL, pp. 197–206 (2005)

An Architecture for Supporting RFID-Enhanced Interactions in Digital Libraries

George Buchanan¹ and Jennifer Pearson²

¹ Centre for HCI Design, City University, London, UK

² Future Interaction Technologies Laboratory, Swansea University, Swansea, UK
george.buchanan.1@city.ac.uk, csjen@swan.ac.uk

Abstract. In this paper, we report the design of an RFID sensing infrastructure for digital libraries. In addition to the architecture of the system, we report its deployment in three different applications to illustrate its use and integration with not only the core DL software, but also web browsers and software for reading documents (e.g. in PDF format). Through this, we demonstrate the utility of RFID support across the entire information seeking cycle.

1 Introduction and Motivation

Digital libraries have tended to live a life detached from the physical world, except insofar as documents are regularly printed out for reading. There has been minimal connection with physical libraries and their infrastructure of books and organised spaces. We believe that this misses a significant opportunity for delivering truly exceptional information seeking support, as much of a user’s information work is conducted in the “real world”. Previous work has focussed on specific parts of the information seeking cycle. One example is the use of digital ink technology, where writing on physical paper is tracked using a specialised pen, and this in turn can be added to digital copies of a document [10]. However, this requires the paper used for printing to be specially prepared, and for the computer to know what logical content is on each individual page.

We introduce a method for using RFID (radio frequency identification) technology, that provides the ability to link physical items with content in a digital library. Our approach complements the previous work on physical interaction with DLs, as it permits interaction with printed books, connects digital notes with physical items, and also enables a range of interactions with catalogs of physical and digital documents. RFID technology is becoming increasingly commonplace, affordable and sophisticated. Whilst RFID is used in libraries to track stock items and to provide security, we believe that there are a host of other uses that are more focussed on the information seeking work that is central to the role of a library. Thus, using RFID in more information- and user-centred ways can unleash new forms of engagement with library users, and better support for the search for knowledge that is the core function of any library. We have investigated this possibility, through a simple componentised service that is built on open-source components, is cross-platform and DL agnostic.

The paper starts by discussing the basic architecture for our system, before continuing to describe one use of the system in detail, and supplementing that with two further example exploitations of the same software. We then discuss the implications of our research, and finish the paper by summarising our main findings and suggesting future avenues for further research.

2 Architecture

In Figure 1 we see the architecture of our RFID support for digital libraries. Presented in the diagram is a traditional library system (centre), that mediates access to external catalogues (right) and delivers content to clients computers operated by users (left). We presume that both physical and digital items in the library are organised by the same topical hierarchy. If physical and digital items are catalogued separately, that does not complicate affairs, provided the same organisational principles are used in each.

In addition to the traditional web browser, the client computer has a second piece of software running on it. This software, the “context capture” client, connects to any RFID reader on the computer, and collects from the sensor, data on visible RFID tags, etc. On the right is the library catalogue, with the traditional core server presented at the top. When the user requests a web page from the server, ①, their web browser connects with the user’s server in the entirely normal model for any web-based application (or web-based DL). This server is supported by the context service. When a web request is received, it contacts the context service ②, which then connects to the context capture

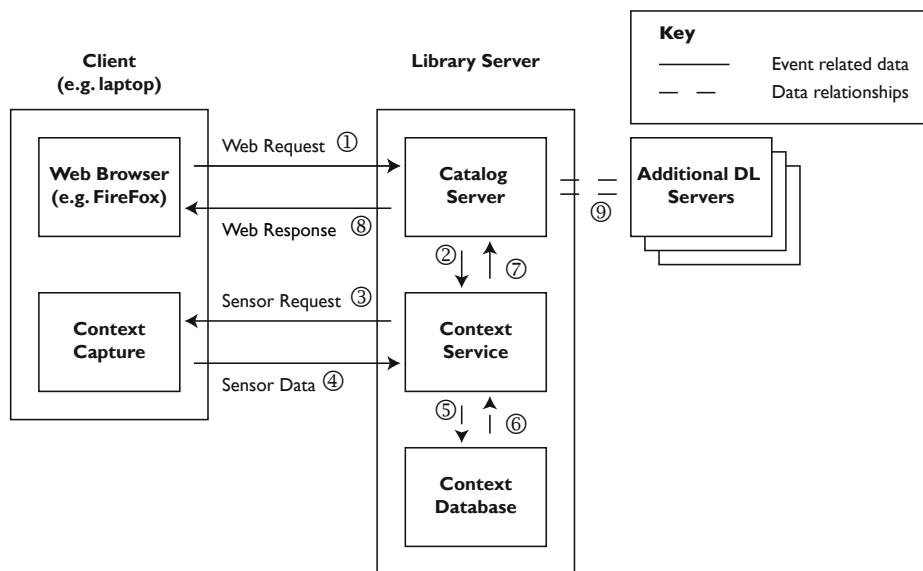


Fig. 1. General architecture of the RFID system

client on the user’s computer ③ to obtain the current sensor data ④. This data is converted from raw ID numbers etc. into known library assets by reference to the context database (bottom) ⑤. Tags that correspond to known assets result in the return of the identity of that asset to the context server ⑥, and then to the library server ⑦: e.g. a tagged book will return the library’s unique identifier for that book. Finally, any adjustments made as a result of the context are made and the resulting response is sent back to the user, again as per a normal web process ⑧. Further information may, optionally, be routed to the user using external resources (e.g. identifying relevant external collections, ⑨).

We will demonstrate an alternative architecture later in the paper, using the same context support software, but in a different configuration. The order of communication here is specific to this particular configuration.

2.1 Implementation

For the implementation of the client side of our system, we use the LibNFC library as the foundation for our hardware-level interface. LibNFC is compilable for all common platforms (including MacOS, Windows and Linux). This core is then wrapped in a Java Native Interface holder to abstract the underlying RFID library and provide a simple cross-platform API. The server component uses the Java Database Connectivity (JDBC) API to store and retrieve relationships between the physical sensor data and library items such as books. Internal communication between client and server modules uses a simple TCP/IP based protocol using the REST paradigm. We use the same protocol method for DL server communication, so a DL server that uses the RFID sensor information only needs to communicate with a platform-independent protocol.

The request used by the context server to obtain tag information from the context capture client is very simple, as it has no parameters. The client returns an XML-formatted list of RFID identifiers recently seen by the computer (the exact details will be covered shortly). As we use a full RFID library, other messages can store additional information onto an RFID tag (tags often contain a small amount of storage). This requires the use of the second function in the protocol, which contains four parameters: the RFID tag to write to, its access key for writing, the place to record the data, and the data block to be recorded. Again, this will be addressed later in the paper. We term our current prototype “EmLi”, or “Embedded Library”. N.B. We use the term “context”, as in fact our implementation supports both RFID sensing and the use of Bluetooth for short-range location identification (which is outside the scope of this paper).

3 Context from Printed Books: Directing Searching

We now demonstrate the use of the EmLi RFID middleware to support contextual information seeking in a digital library. We use the Greenstone DL system [13] as the example DL, and utilise the componentised DL architecture advocated by Suleman and Fox [12]. The RFID sensing facilities are used to adjust the interaction between the user and the digital library catalogue.

Normally, when interacting with a digital library, the information that determines how the library operates is typically entered by the user. Whilst some information can be saved (e.g. preferences for ranking or the presentation of search results), the critical inputs to direct the library’s actions are more often interactively entered (e.g. by the user typing in some query terms and hitting a “search” button). Our previous work successfully demonstrated that a user’s organisation of chosen digital documents can be used to enhance the library’s interaction with the user [5]. One example improvement is to filter search results, in order to highlight material similar to a specific group of documents. However, this previous work was limited to the user’s work to the digital domain, and many information seekers need to work with both printed and digital documents. Hence, we apply the same concept to the use of physical books.

We therefore explored a means for supporting the tailoring of a user’s interaction with their digital library through the printed documents that they are working on. The user can ‘scan’ a book, and the identity of the book is now to be used to filter the library’s interaction. The basic interaction is controlled directly by the architecture already reported in Figure 1. In our current implementation, a USB connected RFID reader is monitored by our Java-based context client running on the user’s computer (a laptop, say, or library catalogue console). This client captures the identity of tags that are sensed by the RFID reader. A list of tags is maintained for a single user session. This RFID data can be flushed if the digital library identifies a change of user – e.g. by a logout – or after a set period of time. However, we will focus upon a single user session for the purposes of this paper.

There are physical constraints that influence our current implementation. RFID readers can only detect tags at a short range (c. 2-5cm for a “high frequency” RFID reader, up to 2m for a “ultra-high frequency reader”). In the former case, this means that a tag is likely only to be detectable for a short period (when a user holds a tagged book next to the reader), and thus we cache book information for two hours. This information is all collated on the client software, and the library only becomes aware of it when interaction occurs with the digital library. When the library server receives a web request (e.g. for a search), it calls the RFID capture software through the sequence described in Sec. 2 to receive a list of known library items. Typically, this is a book, but a tag can also simply represent a node in the subject classification hierarchy (something that, like a document, usually has a unique identifier in a DL [1]). This information is collated to build a “profile” of the common topics of recently scanned books (or topics). In turn, this is used to focus the results given to the user. In a naive application (as is the case with our prototype) this simply uses more classifications to restrict the results). We will now describe this process in greater detail.

3.1 Greenstone and Modularisation

Before describing the use of the RFID components within our pilot, we first need to briefly recap on the standard structure of most DL software. Though

we will focus on the specifics of Greenstone, the same comments can be made of alternatives such as DSpace, Cheshire 2 and Fedora.

Greenstone has two main elements: the ingestion, or “collection building”, process and the run-time interaction with the user. The use of RFID data requires some changes to each process, but the differences are localised in each case. Both parts of Greenstone are modularised, but in different ways, and we shall now discuss the build-time and run-time modifications in turn.

For the build-time, the main change required is to populate the RFID database. To map a detected tag to a particular document, all that is required is a simple link of a document’s unique identifier ([1]) to a particular RFID tag. One approach is simply to add this piece of information to the metadata on a document. As Greenstone is agnostic regarding metadata structures, adding an RFID field to the metadata on a document is straightforward. However, a better separation, can be drawn by placing the RFID data in a separate database. This means that the bibliographic metadata can be kept clear of the RFID data, which is particularly beneficial where the metadata scheme cannot encode the RFID data (e.g. due to limitations of a standard scheme). This approach – which we have implemented – adds a plugin to Greenstone’s ingest configuration, through a new “RFIDPlug” plugin [4]. This plugin simply looks for XML files that follow a particular DTD (for the RFID data required). This is then stored in a separate RFID database than connects known RFID items to specific document IDs. As already noted, document IDs are commonplace features of DL systems, and this method will readily be utilised in any standard DL software.

Turning now to the run-time support of RFID identification, this requires a more extensive set of changes to the Greenstone architecture. Greenstone’s runtime software operates as a web CGI script (in the case of Greenstone 2) or as a Java Servlet (in the case of Greenstone 3). The basic architecture is the same in each case, though our initial implementation is built in Greenstone 2’s C codebase. All web requests are routed through a central web application (i.e. a binary executable for Greenstone 2), which then delegates the request to a specific *Action* component depending on the type of request. For example, browsing a hierarchy results in the *BrowseAction* component being used, whereas a search calls the *QueryAction* component. Extending Greenstone to support the RFID identification requires two changes: first, capturing the RFID data from the client and translating this into corresponding library items (e.g. documents or classifications); and second, utilising that data in the different *actions* as appropriate. We will describe the translation of RFID tags into library items shortly, but we first describe the impact on the Greenstone run-time.

The first change made was to adjust the “receptionist” code that receives web requests and then dispatches the request to the pertinent Action component. The receptionist code was extended to call RFID client component to obtain a raw list of RFID tags. This list is then immediately sent to the RFID server database (see Figure [1]), which then returns a document identifier (for book tags) or classification identifier (for topic-tags) for each RFID tag. This list of library elements was then processed by the RFIDContext module (described below) to

translate this raw data into a set of classifications. These topics are then added to the data sent to the Action module associated with the specific web request.

Many Actions had no adjustments made to them. Two actions were altered: *QueryAction* and *DocumentAction*. The *QueryAction* component was adjusted to use the classifications supplied from the context server: when a search ran, the classifications associated with scanned book and topic tags, were applied as filters to the results, so only documents that matched both the user's query terms *and* the list of classifications were returned. For *DocumentAction*, the topics were used to provide a list of three related documents.

3.2 From Book to Classification

Books are uniquely identified by the RFID tag added to them. This tag is associated with a (catalogue) item in a library. Physical items typically will only have bibliographic data available, and will also be associated with one or more subject classifications in a topical hierarchy. The raw RFID identity consists of a tag type (there are currently about 10 in common use) and the identity itself (unique within each tag type). A simple table in a relational database can note a one-to-one mapping between RFID identity and the document's unique identifier in the library system. As already noted, we optionally permit a tag to directly represent a node in the subject classification scheme, but either through that or the catalogue data on each item, we can construct a list of the subjects identified from different book or topic tags within a user session. This is done by the context server (see Fig. 11). Whilst more complex strategies can be employed our method is currently as follows:

1. If there are one or more common sub-trees for the current set of books, we select those sub-trees for the topic list
2. If there is no common sub-tree, but there are two or more frequent sub-trees (each matching 33% of the detected books), then those sub-trees are chosen
3. If there are no shared sub-trees, then no topic list is built

This is a very crude approach, and we see substantial scope for better models that draw from both theoretical approaches, and also from observation of patterns of real user behaviour. At present, our simple concern is to provide a common "proof of concept" approach that be adopted and improved on by others. When a set of topic is identified, these are then added as constraints when the user searches for information in the digital library.

Once a set of topics has been generated, then the DL system can now utilise that focus in its interaction with the user. For some actions – e.g. reading a document – then we do not currently use this information, though it could be used to create lists of "related books" or other links on the document page.

3.3 User Evaluation

We have installed this EmLi prototype in a pilot project at two UK universities. In a three-phase set of user studies, including observations, interviews and a

traditional pilot study in situ, we have investigated the use of the system, and potential opportunities for using it in future. Initial responses to the twelve person pilot study were positive. No problems were encountered with scanning the tags, and an overall assessment on a five-point likert scale produced one neutral, five positive and six strongly positive responses when comparing the prototype with using the traditional library. Positive aspects included the selection of classifications, where the link between “real” items and the library topics was reported as being easier to understand than lists presented in the traditional DL interface. To quote one participant “I find it hard to make sense of lists, because it takes ages to go through them. I can do that with real books more easily....I don’t feel I lose my place so much.”

The traditional transaction model of web server/browser interaction is not that favourable to providing a slick, continuous interaction. One problem that was encountered was that users expected a very rapid reaction to a tag when scanned: anticipating that the web page would update itself. Five users found the delay frustrating in the current design. The use of AJAX limits the scale of this problem, but the necessary lag of response through web applications is still frustrating. A closer integration between browser and the context client would be helpful, but this is a substantial piece of work in its own right that we have not yet had time to complete outside of using a Java based browser, which makes such extensions trivial to implement.

3.4 Summary

In this section, we introduced the basic alternatives within the framework described in Section 2. The system is relatively simple, and builds on previously proven DL components. When studied in use, the system was favourably received by our users, and a number of suggestions were made.

4 Further Examples

We now demonstrate the use of RFID in two further cases: the connection of printed books to digital notes and documents; and a catalogue interface where the current display can be captured and restored using an RFID card.

4.1 Context from Printed Books: Supporting Notetaking

It is widely recognised that despite the ever growing popularity of digital documents, users often print documents for reading. When taking notes, however, there is often a preference for digital text: e.g. easier editing, text search, and copying. The motivation for this system (see Figure 2) was to allow a simple method of digitally marking-up physical books by providing an electronic annotation space, and connecting the physical and digital media.

Each book in a library is assigned an RFID tag which is used by the system to uniquely identify it. We send a request to ISBNDB.com with the ISBN number to retrieve other information about the book: e.g., the title, author and publisher.

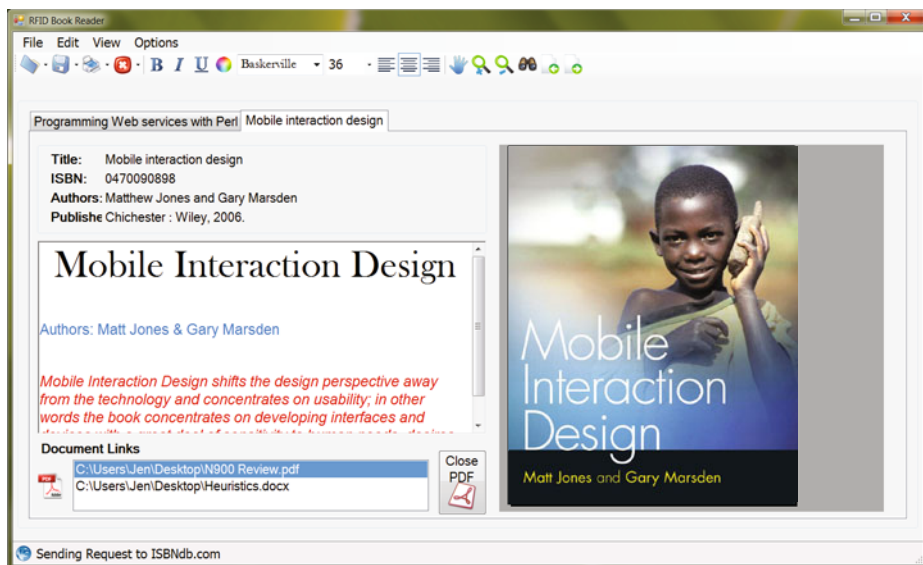


Fig. 2. Screen Shot from the RFID Book Reader System

Once this is complete, the system will display an electronic copy of the document, downloaded to and stored on the local PC. This is shown alongside a rich text editing area (that is saved as .RTF) to make related notes. This method means that users can read from a physical document while conveniently making digital notes that can be easily copied, edited and exchanged. In addition, the electronic copy of the document enables digital tools such as text search and copy/paste which speed up the note-taking process.

As well as making notes within the system, users can also drag other relevant documents into the 'document links' section to keep everything organised neatly. Opening several books simultaneously can be easily achieved by the tabbed menu system, simplifying the research and cross-referencing of multiple documents. If the user does not own a physical copy of a particular book, they can search the database for an electronic copy (that will also have the same edit and link features) by entering: the book title, an ISBN number, the author or any keywords.

4.2 Capturing Catalogue Information

Section 3 showed the connection of an object in a physical library to content in a digital library. However, that only uses the RFID card as a passive tag. We now report the use of the limited storage capacity (64 bytes to 4k) of RFID tags in a user's interaction with a library catalogue. Figure 3 shows the architecture for this application. Compared to Figure 1 the communication sequence is altered. On the left, when an RFID card is presented to the client's reader, the client now actively polls the context server (right) ①. If the detected card is listed

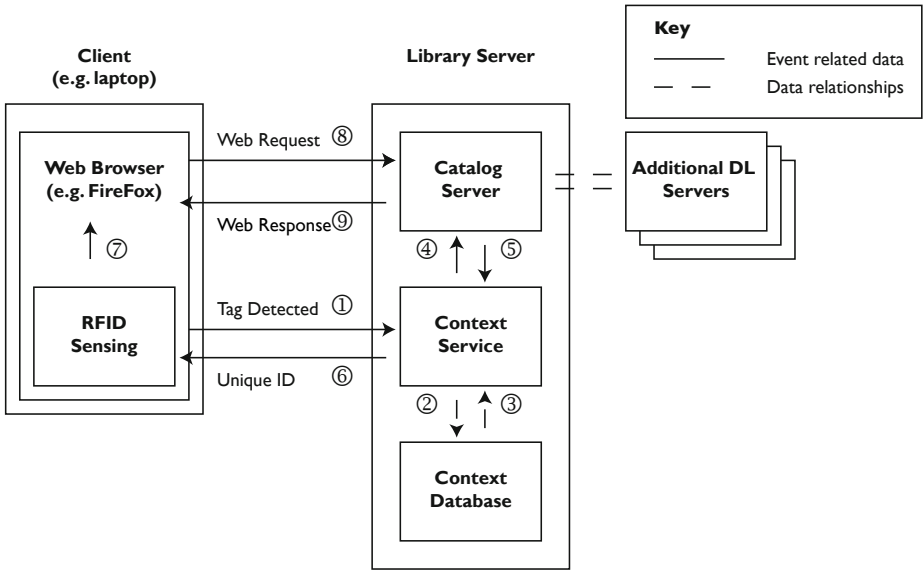


Fig. 3. Modified RFID system, supporting capture from the DL user interface

in the context asset database as a library card (2), then the catalogue capture procedure is initiated (4-9), otherwise we return to the interaction above.

This method requires change to the DL architecture. One method would be to completely integrate the context client with the web browser software, whilst a contrasting approach is to maintain the separation between RFID sensing and the browser (see [8] for a discussion of these alternatives). Regardless of which approach is used, the RFID client software must in this case initiate contact with the DL. The current user page also needs to be captured for later recall. This point is potentially problematic, but for stateless servers like Greenstone, readily achieved through the URL parameters. Where more complex stateholding is used, more adjustment to the DL will be required.

When the RFID client contacts the DL, we record the tag event with a unique identifier on the DL server, and the interaction state is stored with it. The identifier is returned to the client, and the RFID tag records the identifier in its data space. For larger tags, up to 200 identifiers can be stored. The event identifier can be exploited in a number of ways. Previous researchers have adopted a similar method to support physical navigation in a library [11], with an RFID tag used to calculate where the (human) reader should travel to obtain their book. However, we have taken a different approach.

A new *Action* was added to Greenstone - the “stored search” action, that presents a simple list of stored DL states for an RFID card. We only support queries and documents, to capture searches and document matches. This “action” initially presents a page asking the users to present their RFID card. When the reader “swipes” their tag, the context client triggers a second interaction:

rather than a new identifier being created, we retrieve the list of stored DL states associated with the identity of the RFID tag just presented. This causes a page refresh (or, rather, a partial one using AJAX), and the user's list of stored states is presented to them. Thus, a user can "swipe" an RFID tag to capture a particular search - indeed, a particular search page - or a specific document, and then retrieve this through a logical "bookmark" associated with their RFID card.

There are a number of unusual properties of this system that could prove useful. First, the card acts as a key for the saved DL content. This could be used as the basis for privacy support, as it would be straightforward to require both a password and the presence of the card "key". However, this would be unhelpful if the card were lost. It also serves as a means of supporting bookmarking within any point of the DL interface, and is portable between computers. Whilst we acknowledge that some of the features could readily be achieved without RFID use, it is an open question as to what the benefits of using a physical interaction could be to the user. No doubt other researchers will be able to expand and improve upon this current rudimentary implementation.

4.3 Summary

This section described two further RFID applications, implemented on the same basic infrastructure described earlier. Only modest adjustment to the digital library architecture was required, due to the benefits of componentisation and modularisation. However, in both these two cases, and in the previous section, some modification to the user's computer system was required. The basic 'lightweight' baseline of a web browser, that is the common foundation of DL systems such as Fedora and Greenstone, cannot fully exploit the use of RFID tags without the addition of further software to the user's machine.

5 Discussion

Previous systems have investigated some uses of RFID in the library or related areas. We already noted the investigation of support for physical navigation in a physical library [11]. Our work is designed to be more generalisable than theirs, as we endeavour to be DL system agnostic, and our core use of RFID information is much more task neutral. It would be straightforward to extend our basic model to support the physical navigation, introducing way-points as a new RFID type (in addition to books and topic tags).

In other areas, researchers have sought to use RFID, or alternative technologies, to tag physical objects for both navigational and informational purposes. One example is the Marble Museum, which used infrared technology to identify a user's proximity to a specific exhibit [6]. This would trigger the presentation of information specific to the exhibit, in a manner similar to our triggering of topic selection, or the recall of notes, as a result of detecting the presentation of a tag. Other explorations of the same theme have used RFID (e.g. [9,2]), but without adding much to the principles underneath. Again, the software has been

very tied to the specific application at hand, and a very limited set of actions (e.g. obtaining information on an exhibit). Further constraints have applied due to a tendency to use handheld computers (e.g. [2][6]) and these often have limited support for multitasking and simplified browsers. This again underlines our distinct approach in endeavouring to provide a generic service.

Our session-based approach to collecting RFID tags is rather simplistic, and could be radically improved. One key avenue for future work would be to use time as a key element of determining changes in topic. The use of time sequence clustering has already been used effectively in digital libraries for the purpose of organising digital photographs [7], and we are currently experimenting with methods that use the same principle with time-clustered sets of document scans – as may occur when a reader collects three or four books from the library stacks and reads them in sequence in a short period of time. Similarly, our rough heuristics for identifying pertinent topics could itself be substantially refined.

Academic understanding of user navigation in a physical library, and their natural interactions with physical books, remains limited. This is one area where further work would very much add to our comprehension of both how to exploit EmLi as is, and also how to extend its functions, perhaps using other sensors, to achieve a closer bond between physical and digital information work. We have already reported some of our initial findings [3], but we consider the progress made to date to be only the starting point.

6 Conclusions and Future Work

We have introduced a basic infrastructure for supporting RFID-enhanced interaction with physical items and digital libraries. This basic method has already been proven in use in a simple pilot project. We demonstrated the utility of the approach in three separate applications, and the integration of the componentised RFID support into a common, mainstream DL system. Throughout, we have highlighted alternative designs that we have not had the opportunity to fully explore yet, and we hope that these will spur other researchers into building on and surpassing the prototypes we have produced so far.

In future, other sensor data can be gathered using the same model. We have already extended the RFID support to also include Bluetooth facilities, for supporting the identification of a user's location, and hence potential topical interests, when in a physical library. However, this is an area of rapid innovation, and thus no doubt new types will soon become available. We already plan to extend the core functionality with additional features, e.g. for assisting physical navigation, and there is a great need to better understand information seeking as a physical activity. Thus, there is an ample supply for challenges, both technical and regarding human-computer interaction, for future work.

Acknowledgments

This research is supported by EPSRC Grant EP/F041217. Jennifer Pearson is supported by Microsoft Research.

References

1. Bainbridge, D., Buchanan, G., McPherson, J., Jones, S., Mahoui, A., Witten, I.H.: Greenstone: A platform for distributed digital library applications. In: Constantinopoulos, P., Sølvsberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 137–148. Springer, Heidelberg (2001)
2. Boehner, K., Gay, G., Larkin, C.: Drawing evaluation into design for mobile computing: a case study of the renwick gallery's hand held education project. *International Journal on Digital Libraries* 5(3), 219–230 (2005)
3. Buchanan, G.: The fused library: Integrating digital and physical libraries with location-aware sensors. page (in press, 2010)
4. Buchanan, G., Bainbridge, D., Don, K.J., Witten, I.H.: A new framework for building digital library collections. In: JCDL 2005: Procs. 5th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 23–31. ACM Press, New York (2005)
5. Buchanan, G., Blandford, A., Thimbleby, H., Jones, M.: Integrating information seeking and structuring: exploring the role of spatial hypertext in a digital library. In: Procs. ACM Conf. on Hypertext and Hypermedia, pp. 225–234. ACM, New York (2004)
6. Ciavarella, C., Paterno, F.: The design of a handheld, location-aware guide for indoor environments. *Personal and Ubiquitous Computing* 8(2), 82–91 (2004)
7. Graham, A., Garcia-Molina, H., Paepcke, A., Winograd, T.: Time as essence for photo browsing through personal digital libraries. In: Procs. ACM/IEEE-CS Joint Conf. on Digital Libraries, pp. 326–335. ACM, New York (2002)
8. Karpischek, S., Magagna, F., Michahelles, F., Sutanto, J., Fleisch, E.: Towards location-aware mobile web browsers. In: MUM 2009: Procs. of the 8th Int. Conf. on Mobile and Ubiquitous Multimedia, pp. 1–4. ACM, New York (2009)
9. Mantyjarvi, J., Paterno, F., Salvador, Z., Santoro, C.: Scan and tilt: towards natural interaction for mobile museum guides. In: Procs. MobileHCI 2006, pp. 191–194. ACM, New York (2006)
10. Norrie, M.C., Signer, B., Weibel, N.: Interactive paper as a reading medium in digital libraries. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 232–243. Springer, Heidelberg (2008)
11. Satpathy, L., Mathew, A.P.: Rfid assistance system for faster book search in public libraries. In: CHI 2006 extended abstracts, pp. 1289–1294. ACM, New York (2006)
12. Suleman, H., Fox, E.A.: Designing protocols in support of digital library componentization. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, pp. 568–582. Springer, Heidelberg (2002)
13. Witten, I.H., Boddie, S.J., Bainbridge, D., McNab, R.J.: Greenstone: a comprehensive open-source digital library software system. In: Proc. ACM Conf. on Digital Libraries, pp. 113–121. ACM, New York (2000)

New Evidence on the Interoperability of Information Systems within UK Universities

Kathleen Menzies, Duncan Birrell, and Gordon Dunsire

Centre for Digital Library Research, University of Strathclyde,
Livingstone Tower, Richmond Street, Glasgow, G1 1XH
klmenzies@cis.strath.ac.uk, duncan.birrell@strath.ac.uk,
g.dunsire@strath.ac.uk

Abstract. This paper will report on the key findings and implications of the JISC-funded Online Catalogue and Repository Interoperability Study (OCRIS), a 3 month project which investigated the interoperability of Online Public Access Catalogues (OPACs) and Institutional Repositories (IRs) within UK Higher Education Institutions (HEIs). The aims and objectives of the project included: surveying the extent to which repository content is in scope for OPACs and the extent to which it is already recorded there; listing the various services to managers, researchers, teachers and learners offered by these systems; identifying the potential for improvements in the links from repositories and/or OPACs to other institutional services such as finance or research administration.

The project combined quantitative and qualitative methods; primarily, an online questionnaire distributed to staff within 85 UK HEIs, purposive sampling and two in-depth case studies conducted at the Universities of Cambridge and Glasgow.

Keywords: Interoperability, digital libraries, repositories, catalogues, standards, resource discovery platforms.

1 Introduction

Although the role of the Online Public Access Catalogue (OPAC) as a user-facing front end offering access to bibliographic records contained in the Library Management System (LMS), is well-established, the emergence of Institutional Repositories (IRs) signals the arrival of competition for this function.

OCRIS - The Online Catalogue and Repository Interoperability Study [1] - investigated the interoperability of OPACs and IRs within UK Higher Education Institutions (HEIs). Changing user requirements and web technologies ensure that both types of system are expected to be flexible and adaptable in response to trends such as 'Web 2.0' and the Semantic Web. The stark economic and competitive pressures facing Universities add further imperatives such as the need to develop value-added services useful to a range of stakeholders.

Eighty-five UK Universities have at least one IR - with more planned for many of the remainder. All have an OPAC [2]. Universities planning to further develop

¹ This figure was correct in mid-2009, when the project began.

Digital Library services must consider both systems to be important components. Both can be located within a wider Information Systems environment which includes Research Management Systems (RMSs), Publications Management Systems (PMSs), Current Research Information Systems (CRISs), Virtual Research Environments (VREs) and many other types of administrative database (such as those used by Finance or Human Resources departments). Whether or not the theoretical relationship between these systems is being actualised through concrete implementations depends, at a high level, on the policies of the parent institution and - at a lower one - on the extent to which they interoperate.

There is little consensus about the ways in which LMSs, OPACs and IRs might be developed to support internal and external processes and workflows [2,3]. This may be partly due, as Lagoze convincingly argues, to a failure by institutions to recognise the complex impact of openly accessible online information and “the attempt [by Digital Library developers, largely characterised by individual projects rather than by agile, innovative communities] to deploy an information infrastructure that essentially retrofit[s] new technology on traditional information models” [4].

If key components of Information Systems cannot communicate; if there are no interfaces between them; when syntactic and semantic interoperability remain unsupported, systems will be ineffective in meeting the information seeking and data assembly requirements of many end users.

The absence of common definitions and an inability to resolve differences between (or logically group) distinct vocabularies used by information systems, frustrates attempts to share data and to develop a scalable Digital Library. For instance: a federated architecture utilising shared namespaces and unique identifiers would allow a grid of machines to query the same information pool [11], yet this would require new and clearly focussed policies and workflows.

At present, a lack of interoperability significantly reduces the likelihood of institutions becoming the collaborative, active, networked environments that are now essential for process efficiency within modern organisations aspiring to high technology [5].

Strong supporting evidence for this comes from the pilot data collection exercise carried out in 2008 by 22 UK Universities in preparation for the proposed bibliometric aspect of the Research Excellence Framework (REF). For many, this proved extremely problematic due to the variety of places where data were stored - a “broad spread” of decentralised systems (including paper-based ones) that differed in scope, sophistication and data quality [6].

Given the well-understood advantages of interoperability (and in a climate of immense sector-wide financial strain) it is worthwhile assessing whether library and other systems currently constitute a network in any functional sense within UK Universities. Is the existing landscape characterised by strategic, streamlined thinking or by fragmentation? What examples of good practice might be identified? Over the three month period allotted to the project, OCRIS set out to answer these questions.

2 Methodology

A combination of qualitative and quantitative methods was used by the project team. Taken together, the gathered data would constitute a representative survey of all 85 in-scope HEIs.

Three questionnaires were constructed and made available online. One was aimed at IR managers and developers, one at systems librarians and bibliographic services staff, and the third at those working with non-library administrative systems. This third questionnaire was comparatively brief due to the wide-ranging and usually bespoke nature of systems falling under the label “administrative”.

Recipients of the first two questionnaires were asked about the content types held by their systems; the standards and authority controls in place within each; compliance or otherwise with data harvesting protocols; resource discovery tools and integration services being used or considered; the other institutional systems with which theirs do or do not interoperate, and the extent to which metadata entry is mediated by staff. Using an ontology based largely on the one proposed in Linking UK Repositories [\[7\]](#) - a detailed report which identifies those services most necessary for the creation of an effective repository network - the second half of the form asked respondents which services (if any) their systems currently support. Space was also provided for comments.

Over one thousand individuals were sent links to the OCRIS questionnaires. Two-hundred and sixty two of these were IR managers or staff, 280 worked within library systems teams or bibliographic services departments and 599 worked within finance, research co-ordination, human resources, auditing, quality assurance and information systems and services departments.

In parallel to the questionnaire dissemination, a workflow was devised by which to study in-depth the OPACs and IRs within 10 HEIs deemed (both individually and as a group) to be illustrative of a particular facet of the project and its context. This purposive sampling provided a snapshot of the current situation regarding duplication of records and scope between bibliographic/publication systems, the way item types are described and curated, the implementation and customisation of the software being used and the extent to which links are being made between the two systems.

Two case studies were carried out within the libraries of large, research-led Universities. By discussing the aims of the project with staff at Cambridge University Library and Glasgow University Library, specific issues pertaining to systems development, the selection of technologies, the formation of policies and workflows and the concerns of those working with library systems on a daily basis, were made clearer to the project team, with examples of good practice identified.

3 Key Findings

The response to the OCRIS questionnaire from the Institutional Repository community was satisfactory, with data gathered from 41 percent of all institutions

contacted (i.e. 31 institutions out of 85). Responses from the other two groups were far fewer (18.2 percent of institutions (16 out of 85) responded regarding LMSs and only 8.2 percent (seven institutions) regarding administrative systems). Because many institutions have more than one IR or OPAC, the total number of these systems represented in the responses is 58 (rather than 47).

It seems legitimate to speculate that at present, the IR community is most active in terms of system development and experimentation; hence their greater level of engagement with the project. Whereas the LMSs used within UK HE are entirely proprietary, development may become more possible when systems make use of Open Source software (as all IRs included in the OCRIS questionnaire do). However, factors affecting response rate were not studied.

In some instances, contradictions arose between the responses given by LMS and IR staff working within the same institution when asked whether their systems interoperate². This complicated the process of analysis although, as indicated in Figure 1 below, conflicting data were accounted for.

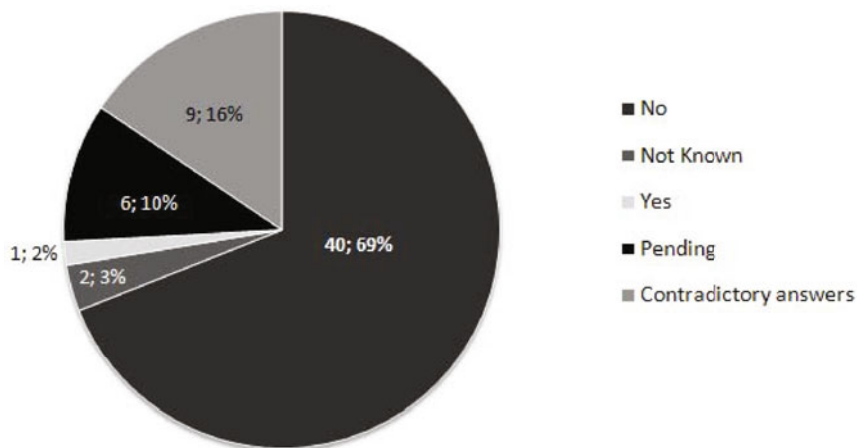


Fig. 1. Responses to the question of whether IRs and LMSs currently interoperate

Despite its insufficiency for definitive statistical analysis, a number of useful comparisons and insights were produced from the OCRIS questionnaire data. These, and key findings derived from the other analyses undertaken, are summarised below.

² While taking into account the complexity of interoperability and its many technical aspects, a simplified definition was used to make it easier for participants to respond to the OCRIS questionnaire: “Generally, interoperability refers to how well two or more systems work together to achieve a common goal; here, this means direct processing by one automated system or sub-system, of data provided by another. This will usually be assisted by the use of standards and standard protocols”.

Table 1. Other institutional systems with which IRs and LMSs share data

System Type	IRs interoperating with	LMSs interoperating with
None	6	11
Electronic Resource Management	1	7
Human Resources	7	6
Virtual Learning Environment	5	6
Finance	0	6
Serials Management	0	6
Other	8	5
Course Administration	0	4
Research Administration	9	0
Research Assessment	8	0
Virtual Research Environment	2	0
Enterprise	1	0

3.1 Interoperability, Services and Data Sharing of IRs and LMSs

The majority of IRs and LMSs within UK Universities do not currently interoperate. Across the group, both the IR and LMS respondents from only one institution (i.e. 2 percent) answered in the affirmative. Even if the responses of those who answered that interoperability was “pending” are added (and even given margin for error), the total remains very low at 8 percent (approximately seven sets of institutional library systems).

The range of other institutional systems with which IRs and LMSs interoperate is higher and is shown in Table 1.

Thirteen institutions (81 per cent) responding to the LMS questionnaire stated that their system currently interoperates with at least one internal system. Of those: six interoperate with four other types of system, three with three, one with two and three with one system only.

Twenty IRs (64.5 percent of institutions) interoperate with other, internal systems. Of those, seven interoperate with only one other system, six with two, six with three and one with four. It is notable that systems concerned with research assessment or management are those with which IR developers most frequently instantiate active links. Although no weighting against the total number of systems within responding institutions was calculated, it is likely that the figures shown above would be very low, given the large number of information systems in place within any given University.

3.2 Administrative Systems

When asked, “Do your systems use Open Standards” two respondents answered yes, one no and the remaining four, unknown. Three said their system did not interoperate with the LMS, two that it did, and two unknown. While three systems interoperate with the IR, with one pending, one answered no and two, unknown.

No respondents stated that they found improving interoperability to be an issue of no importance, yet there was some uncertainty. Four answered yes when asked “Is improving interoperability important?” yet three selected not known. Qualitative data gathered indicates something about attitudes in some instances. Of those respondents who are considering interoperable, integrated research systems, several left comments for the project team. One stated that:

“We currently have an in-house Research Expertise system used for research admin/assessment which links to our IR and other corporate systems as indicated. We are about to... [...]...replace this with a proprietary system using CERIF as the core data model and CERIF-XML³ as the exchange format between systems. We harvest data from third party sources such as ISI and PubMed for journals and similar output; but we manually input metadata on books; so what would be useful would be authoritative source(s) of metadata for non journal items that could be harvested in CERIF-XML; similarly it would be useful to have access to authoritative lists of journals and publishers in this way.”

Plans to increase interoperability (and knowledge of its potential impact) are being demonstrated in some instances. This respondent makes clear the importance of standards (in this case CERIF) and common encoding methods (XML) for information exchange. It is worth asking whether the LMS - a rich source of authoritative metadata pertaining to monographs or books published by staff - will be leveraged, without externally-held records being seen as the primary source.

3.3 Services

The lack of interoperability suggested above inhibits the development of services in support of core activities - for example, the compilation of management reports or other types of statistical account and the generation of publications lists. It also makes finding resources difficult for the end user.

It is particularly noteworthy that for both LMSs and IRs, manual metadata creation and enhancement services (63 and 85 percent, respectively) are the second most frequent type of service currently supported after the generation of usage statistics (73 and 77 percent). Reporting services for specific administrative departments is uncommon but twice as frequent in IRs as LMSs (20 percent, 10 percent). Providing advice on IPR (77 percent) and Open Access is, as might be expected, extremely common (94 percent) within IRs. This seems to suggest that IRs are still in their infancy, with explanations of the basic concepts underpinning them still very much a necessity.

Similarly, only a small number of LMSs are beginning to incorporate services which go beyond what might be considered their traditional role - such as digitisation (12.5 percent of respondents) and the verification of digital objects (25

³ Common European Research Information Format: a European Union Recommendation to member states intended to facilitate data exchange and the resolution of schema differences between heterogeneous distributed databases - primarily those of Current Research Information Systems (CRISs).

percent). Basic services, such as arranging for the provision of adequate, high-quality metadata and monitoring performance remain more significant parts of the workflow within both IRs and LMSs than do other added-value services.

There is little to suggest that the majority of systems used in UK Universities are being developed with a strong, policy-driven emphasis on shared services and processes. Rather, it seems that service provision is somewhat piecemeal and, in general, is fragmented.

3.4 Semantic Equivalence in Description of Scope

Item types (variously termed “Format”, “Medium” or “Material Type”) are usually only viewable via “Advanced Search” menus and are generally derived from fields inbuilt into the system software (e.g. MARC or Dublin Core encoding). In LMSs these fields are often modified to include administrative information relating to circulation control or item location. This means there is little commonality in how items are described across systems and scope is difficult to accurately discern, especially as it may extend beyond the items currently catalogued.

At Cambridge University Library (CUL), for example, items searchable in the OPAC are given as Book, Serial, Electronic Journal, Electronic resource, Disc (CD/DVD), Music Score, Map, Non-Musical Recording, Musical Recording, Archive/Manuscript, Kit, Mixed Material/Collection, Mixed Material, Visual Material. But naturally the true scope extends beyond this list of 14 and is more complex, with varying levels of granularity.

Different pages of the website reveal what is recorded in Newton (the CUL library catalogue) in more detail (for example “maps catalogued since August 2000”, “all printed books published from 1978 onwards, with the exception of Official Publications”). The website also explains that Manuscripts and Theses are recorded in a separate catalogue. The necessity for these additional notes illustrates the complex landscape within which the OPAC is situated and the fractured experience facing end users who search across resource sets for data.

The Cambridge Institutional Repository (DSpace) allows searching by item type: Article, Audio, Book or Book chapter, BW Image, Colour Image, Dataset, Drawn image, Image, Journal Article - Published Version, Journal Article - Submitted Version, Map, Other, Preprint, Presentation, Software, Table, Technical Report, Thesis. These do not correspond well to the item types held by the OPAC.

Yet OCRIS questionnaire data suggests that *every* item type in scope for the OPACs of UK Universities is also now in scope for IRs. Further, boundaries are becoming increasingly blurred, with many IRs containing bibliographic data and OPACs linking to full text. Thus instances of both partial and full scope overlap are significant.

A lack of shared vocabularies or standards means that there is little correspondence between item/material/format types listed in OPACs and IRs; when searching both systems independently it can be confusing trying to discern what the commonalities and differences are.

If standard fields deriving from cataloguing schemes or repository software are used, this can assist in keeping simple the construction of search/browse lists, allowing search/advanced search limiters to be implemented with minimal effort. Here, as long as different terms used by different systems are referred to in the same way by the underlying code (for example, a numeric ID) some amount of data sharing might be made easier (although of course, this is not the same as interoperability).

Describing item types in more detail, to reflect the collections of specific libraries or repositories, is clearly beneficial to end users, but if basic item types are not the same, there is no direct matching and this may complicate the search experience. Consistency and clarity for end users searching both within and across institutional systems would, at present, require resource-intensive mapping or conversion exercises.

3.5 Resource Discovery Platforms

Integration between IRs and LMSs is generally being built from scratch into new “vertical search” products, often referred to as Resource Discovery Platforms (RDPs). This is part of a ‘de-coupling’ of systems, with a focus on the end user; RDPs, which pull in data via standard interfaces, can be placed ‘on top’ of the LMS, supplementing the traditional OPAC. The interoperability/integration services of some well known examples (such as Encore, Primo, AquaBrowser, Talis Platform and VuFind) include OAI-PMH harvesting, SRU/SRW, RSS feeds and the use of web services.

Table 2. Integration services are moderately popular and can facilitate data sharing

Integration Service	Instances within LMSs	Instances within IRs
Metasearch/linking	13	7
Web Service/API	7	10
Vertical searching	4	4
Software as a Service	3	1
Other	2	4

These, rather than the traditional core LMS or its technologies (such as Z39.50), are being used to push forward interoperability and distributed search. RDPs are a good first step towards showcasing the advantages of data sharing. Through import mappings, they translate data at ingest from its representation within each contributing system to the model of which the host system makes use, thereby matching bibliographic fields to create a single record format from the various standards pulled in via its various interfaces.

Although currently being marketed heavily at the LMS community by vendors, not many respondents overall are using or planning to provide vertical search across resource sets (only 17 percent of the total institutions responding) at present. Nevertheless, it can be seen that this and other data linking technologies are being considered by both LMS staff and IR developers.

4 Case Studies

Examples of good practice were discerned from two brief but intensive case studies undertaken at Cambridge University Library and the University of Glasgow Library. At both institutions, the procurement of RDPs has been motivated not only by a desire to address the expectations of users but also to include and expose repository content and other types of learning, research and teaching material within the OPAC. At Cambridge staff felt they were “solving a number of problems - not least the multiple catalogue silo situation.” The library issued an Invitation to Tender for a new RDP in April of 2009, with the product required to:

“...act as a single point of discovery for University collections...It is anticipated that it would also include library materials from digital collections within Cambridge and subscribed electronic material, via a Federated Search service, and by harvesting data from non-bibliographic sources, such as the DSpace Institutional Repository.”

Among other specifications, the platform would need to support, “full harvesting and storage of bibliographic and other forms of metadata from a variety of sources, including the Library Management System” and “a means to control de-duplication of data from incoming streams”. Systems staff provided detailed specifics about how de-duplication should be carried out to ensure that groups of duplicates will be adequately identified and edited/merged to create “best records” to which other incomplete (and hidden) records could be linked. [8](#)

Glasgow has already installed an RDP to act on its library catalogues and is considering how records from its central IR (Enlighten) could be integrated into this. Both Universities are heavily involved in Identity Management activities.

At Glasgow, the University’s Identity Management infrastructure is seen as beneficial to mapping resources against one another through common identifiers. UGL staff viewed the application of consistent, institution-wide “digital identities” known as Glasgow Unique Identifiers (GUIDs) administered by the University’s “Identity Vault”, as providing the means for the greater co-ordination of systems and an increased convergence of strategic services.

The GUID is used by university staff for a variety of services including access to HR and Research Systems and the submission of Time Allocation Schedules (TASs) as part of mandatory reporting activities. Within the library, the GUID is already tied into a number of services and is used to authenticate access to e-books, e-journals and subscription databases.

Cambridge’s “Identity Management Group” (of which members of CUL’s Electronic Services and Systems team are a part) is investigating how HR records relating to staff and students might be better shared across systems, with possibilities including a portable ID that could track an individual throughout their time at the University. This ID would be persistent, being retained even were an individual to graduate and return at some point as a conference delegate or visiting academic.

Implementation of such a system for identity management would be feasible as there are already both University Cards and Library Cards which use the same number for identification purposes. However, there are multiple “data owners” within Cambridge, which has a complex, federated structure. Devising a comprehensive ID or name list from the HR system would be complicated as there is no common login scheme in place. College libraries may use their own ID systems on College library cards and many individuals who do not qualify for an institutional email address are still allowed access to Cambridge libraries. In terms of publications, there may not be any direct link between a named author and the institution; they may have contributed to a paper or symposium but no name authority would exist for them anywhere at Cambridge.

Questions remain, then, as to whether such IDs have sufficient metadata attached to support disambiguation and provide an adequate basis for service development. In terms of the wider context (i.e. beyond institutional boundaries), these developments do not necessarily make the sharing of records with external systems possible; however they do provide a clear potential solution to the problems of data sharing within individual HEIs. Staff stated that local solutions were necessary as they cannot afford to “wait and see” which external name authority projects (e.g. Open Researcher and Contributor ID (ORCID) or Names) will succeed, however promising they may appear.

Identifiers created by Identity Management Systems could become the “glue” of a “Junction Box” model, wherein the IR (or another centralised University system) would become a “portal” for re-usable data linking staff and student records to publications data and a wide range of other administrative information.

In both case study libraries, staff cataloguing items for or developing the IR work closely with those concerned with the LMS and its catalogue records, creating formal and informal shared workflows. This, and the wider discussions with which library systems staff are involved, should be seen as exemplary by other HEIs considering process efficiency, data sharing, and a move towards interoperability.

5 Discussion and Recommendations

Many institutions now possess a range of discipline, department, content or format specific repositories, serving a wide range of user communities with different needs and requirements. Roadmaps for further development must recognise the benefits that the diversity of local systems can bring, including the various ways in which content and metadata might be stored, used, presented and shared, particularly where content and carrier can be usefully separated.

As Jerome McDonough points out [9], “we will never all use the same metadata standard. We need to stop thinking about interoperability as a problem of adopting a single standard, and start thinking about it as a problem of translation.” The key is making links between digital resources by sharing data between systems.

OCRIS found little to suggest that this is currently a strong consideration within the majority of UK Universities, certainly in terms of actual implementation. The range of information systems within HEIs are not being developed with common goals or standards in mind or under common policy. Resource Discovery Platforms are doing something to assist data sharing between library systems; other types of administrative system do not yet appear to be significant parts of the Digital Library.

The needs of both external and internal stakeholders could eventually be accommodated through the development of shared workflows across departments and the establishment of a central data ‘hub’ where records and metadata can be made reusable across all parts of the institution. If interoperability is indeed the goal, then the systems which support data storage and information services cannot be treated as entirely separate; they should operate with relative autonomy, and Open Standards must be used wherever possible.

A combination of socio-technical approaches and business process modelling [10] might assist this effort, which ultimately depends on common standards, protocols and eventually, semantic interoperability, supporting a potentially huge number of services and processes sector-wide. The key recommendations made by the OCRIS project are as follows:

1. Improve co-ordination between all departments possessing information systems, with support from the top-down, to develop efficient workflows, reduce un-necessary duplication of effort and formalise collaboration.
2. Consider establishing a centralised system and attendant workflows for cross-checking and cleaning metadata that is to be shared between systems to ensure quality, usability and re-usability.
3. Consider establishing shared namespaces where terms can be mapped against a set of reserved words. This allows queries by systems using different terminologies and standards to be resolved, ultimately supporting a federated information systems architecture. Namespaces defining document and user types would need to be considered as a first step.
4. Develop clear policies on the scopes and uses of IRs and OPACs - present these clearly, comprehensively and comprehensibly, to both staff and end users.
5. Expose all possible LMS and IR records for harvesting and linking via distributed/federated/meta search using technical protocols such as OAI-PMH, SRU/SRW or link resolvers, as appropriate.
6. Establish or use existing controlled name authority lists for all staff, making these available to all relevant departments.
7. Develop (or if already place, make consistent use of) persistent institutional or departmental IDs, making these available internally and to other institutions.
8. Person and role information from various institutional systems should be stored and made available to all systems as valuable contextual metadata.
9. Recognise that LMSs are a rich source of bibliographic information on books, monographs and other items authored by staff; this should be leveraged

and made available to all stakeholders by the development of interfaces and common standards to extract and share such data.

The OCRIS project marked a first attempt to gather quantitative and qualitative data from UK Higher Education Institutions pertaining to the processes and practices which can improve data sharing, syntactic and semantic interoperability and policy development, particularly with regard to library information systems. It is hoped that the project's findings constitute a useful basis for future work on these topics.

References

1. Birrell, D., Dunsire, G., Menzies, K.: Online Catalogue and Interoperability Study: Final Report (2009), <http://ie-repository.jisc.ac.uk/430/>
2. For example, see Sero Consulting Ltd., Glenaffric Ltd., and Ken Chad Consulting Ltd.: JISC and SCONUL Library Management Systems Study. An Evaluation and horizon scan of the current library management systems and related systems landscape for UK higher education (2008), <http://www.jisc.ac.uk/media/documents/programmes/resourcediscovery/lmsstudy.pdf>
3. For example, see Sherpa: Distribution of UK Higher Education Repositories (2007), <http://www.sherpa.ac.uk/repositories/>
4. Lagoze, C.J.: Lost Identity: the Assimilation of Digital Libraries into the Web. In: A Dissertation Presented to the Faculty of the Graduate School of Cornell University in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy by Carl Jay Lagoze, February 2010 (2010), <http://www.cs.cornell.edu/lagoze/dissertation/CarlLagoze.pdf>
5. Hobday, M., Davies, A., Prencipe, A.: Systems integration: a core capability of the modern corporation. *Industrial and Corporate Change* 14(6), 1109–1143 (2005)
6. Technopolis. Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) Bibliometrics Pilot Results of the Round One Consultation (2009), <http://www.hefce.ac.uk/research/ref/Biblio/>
7. Swan, A., Awre, C.: Linking UK Repositories. Technical and organisational models to support user-oriented services across institutional and other digital repositories (2006), http://ie-repository.jisc.ac.uk/10/1/Linking_UK_repositories_report.pdf
8. The Chancellors, Masters and Scholars of the University of Cambridge: Invitation to Tender. Tender Document for the supply of a Resource Discovery Platform (2009)
9. McDonough, J.: OAIS, Designated Communities and Metadata (2008), http://www.digitalpreservation.gov/news/events/ndiipp-meetings/ndiipp08/docs/session11_mcdonough.ppt
10. Borbinha, J.: It is Time for the Digital Library to Meet the Enterprise Architecture. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) *ICADL 2007*. LNCS, vol. 4822, pp. 176–185. Springer, Heidelberg (2007)
11. Reilly, S.: Tupelo-Scheck: Digital Object Repository Server: A Component of the Digital Object Architecture. *D-Lib Magazine* 16(1/2) (2010), <http://www.dlib.org/dlib/january10/reilly/01reilly.html>

Enhancing Digital Libraries with Social Navigation: The Case of Ensemble

Peter Brusilovsky¹, Lillian Cassel³, Lois Delcambre², Edward Fox⁴, Richard Furuta⁵,
Daniel D. Garcia⁶, Frank M. Shipman III⁵, Paul Bogen⁵, and Michael Yudelson¹

¹ University of Pittsburgh, PA 15217

² Portland State University, Portland, OR 97207

³ Villanova University, Villanova, PA 19085

⁴ Virginia Tech, Blacksburg, VA 24061

⁵ Texas A&M University, College Station, TX 77843

⁶ University of California, Berkeley, CA 94720-4600

Abstract. A traditional library is a social place, however the social nature of the library is typically lost when the library goes digital. This paper argues social navigation, an important group of social information access techniques, could be used to replicate some social features of traditional libraries and to enhance the user experience. Using the case of Ensemble, a major educational digital library, the paper describes how social navigation could be used to extend digital library portals, how social wisdom can be collected, and how it can be used to guide portal users to valuable resources.

Keywords: social navigation, digital library, portal, navigation support.

1 Introduction

The presence of other patrons enhances library experiences even for those who come to the library alone. Just think about the value of “others” in a university library. Following their peers with similar needs and interests, students can easily locate sections and shelves where relevant books are located. “Wear and tear” of book covers directs them to most used resources. Bookmarked pages and pencil comments allow finding fragments that may be relevant for a specific course within a book (for that reason many students are known to prefer used textbooks). Altogether, the past activities of other patrons, their signs of attention, help future users to locate the most important and interesting books and fragments, which could be overlooked otherwise. Unfortunately, the social nature of the library is typically lost when the library goes digital – just as with many other digital artifacts – Web sites, digital museums, classrooms, etc. The loss of the social side of digital artifacts motivated research on *social navigation* [6], which attempted to collect traces of past users and make them visible in some form to help future users.

We believe that digital libraries (DL) should attempt to replicate the social features of traditional libraries to serve their users better. It was among the main goals of our NSDL Ensemble project to turn an educational digital library into a social place. The

Ensemble portal is engineered to collect various traces of user interaction with the portal. These traces represent the collective wisdom of the various user groups working with Ensemble and the portal attempts to use this wisdom to guide future users of the portal through *social navigation*. Social navigation guides users to useful and interesting resources through adaptive link annotation and link recommendation. This paper uses the Ensemble case to explain how social navigation could be used to extend digital library portals. It explains how social wisdom can be collected by various tools available at the portal and how it can be used to guide portal users.

2 Social Information Access and Social Navigation

Social navigation [6] is a type of *social information access*, a group of techniques that employs users' past interaction with an information system (known as explicit and implicit *feedback*) to provide better access to information to the future users of the system [2]. Different social information access techniques are typically categorized by three aspects: (1) which kind of past user behavior it collects; (2) how these traces are processed to form "community wisdom"; and (3) how this information is used to enhance user information access. Social navigation techniques are similar in the third aspect: they assist the user in the process of navigation (browsing) by adapting links used for navigation. The process of link adaptation may include re-arranging existing links, enhancing links with adaptive visual cues, or generating new links. Different social navigation techniques can be based on different kinds of past user behavior.

The most popular kind of user behavior used for social navigation is user browsing. The social navigation based on browsing behavior (sometimes called as *traffic-based* navigation) has been used in a number of projects [3; 7]. More recent projects attempted to increase the reliability of social navigation by using user annotation behavior [8] and bookmarking [9]. However, social navigation "beyond traffic" is still rare. In this context, the Ensemble project attempts to extend past research on social navigation by exploring a range of user actions as a basis for social navigation.

3 Social Navigation in Ensemble

Ensemble is the Computing Portal in the US National Science Digital Library (NSDL), an ambitious project to provide access to learning materials and resources for education in the Science, Technology, Engineering and Mathematics (STEM) disciplines at all age and education levels. Ensemble provides a distributed portal to such materials, allowing both multiple sources and also multiple access points to those resources. Ensemble explicitly supports the role of a library as more than a repository by emphasizing a role as a gathering place, a place for sharing, for finding others with similar needs and interests, a place for meeting to work on a common project, a place for highlighting people and resources of particular current interest.

The key component of Ensemble is the Web portal: www.computingportal.org. The portal presents easy access to recognized collections and tools. It is also a central meeting place for communities who are interested in various aspects of computing education. The three themes of the portal: collections, communities, and tools represent a social connection among people and between people and resources.

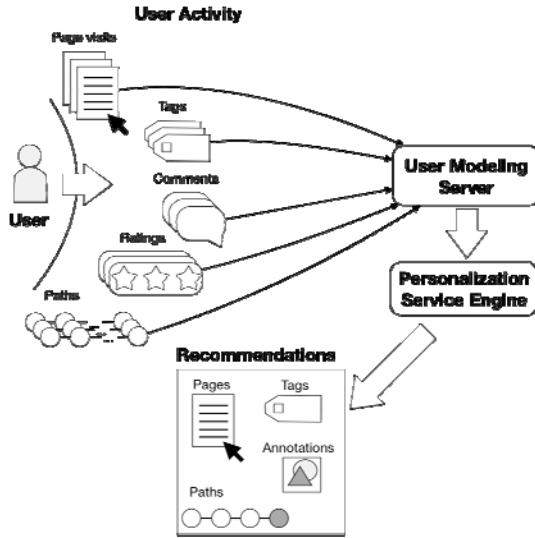


Fig. 1. The social navigation architecture of Ensemble portal. Various traces of user actions are collected and processed by the user modeling server. The accumulated “wisdom” of the user community is used by personalization services to provide social navigation.

The collective wisdom in Ensemble is assembled by tracking various actions of portal users. The Ensemble portal tracks both traditional low-level user actions such as resource browsing, rating, commenting, and tagging; and higher level structural actions such as fragment extraction and composition. The information about all these actions is accumulated in the user modeling server of the portal, which processes this information and makes it available to social navigation services (Fig. 1).

Following the nature of Ensemble as a meeting point for various groups interested in computing education, the Ensemble portal processes the collective wisdom on two levels: the *portal level* and the *group level*. The portal level integrates traces of all portal users, while the group level integrates actions of a specific community or group of users such as, for example, a CS1 community, a group of users focused on the first-year course in computer science (<http://www.computingportal.org/cs1>). Each user working with Ensemble can explore the portal as a member of such a group. In this case, each action of this user contributes to both the portal wisdom and the wisdom of her current group. In turn, an association with a specific group enables group-level social navigation, i.e., exploring the portal as a member of a group, the user is guided by both portal-level and group-level wisdom.

Group association is neither mandatory nor exclusive. A user can explore the portal without a group association (in this case only portal-level wisdom is used). A user can also belong to several groups and communities and change her current group at any time. However, at any given time the user can be dynamically associated with one group and explore the portal from one perspective since social navigation guidance offered by different groups could be contradictory.

4 Collecting Social Wisdom in Ensemble

Ensemble integrates experiences of several earlier social navigation projects and tracks a wider set of user actions than earlier systems. Feedback actions, which are traditionally used in social navigation systems, are focused on a specific resource and provide explicit or implicit feedback on that resource. These actions can be used to estimate user and group level of interest in the resource. Application actions comprise using resources in the context of user needs. The usage is traced on both the level above and below stored resources. The level above means composing artifacts from several resources (such as the Ensemble guided path mechanism). The level below means extracting a fragment of the resource for re-use. Tracking these actions allows Ensemble to support more sophisticated social navigation.

4.1 Tracking User Navigation

Ensemble uses Drupal, a content management system that has gained popularity for its dynamic features. Based on user-id, we can track what pages a user viewed on a particular date. It also lists the hostname, which can be used to track down the geographic origin of a page request. Tracking user browsing outside the Drupal portal is more complicated and can be done by using *embedlets* (see section 5) and Google Analytics. Both kinds of browsing history can be processed by the user modeling server and applied to provide social navigation. For example, link annotation attract user attention to resources, which are most popular among the group users while link generation can recommend resources, which were visited next by users from the same group in a similar context.

4.2 User Direct Feedback: Comments, Ratings, Tagging

A state-of-the-art educational portal nowadays is expected to offer three kinds of user feedback known as CRT (Comments, Ratings, Tagging) [12]. Comments and ratings are examples of explicit feedback, which allows users who have enough interest in a resource or a post to augment the content in some way.

Comments are the most extensive. The user contributes something to the item by extending the information provided or by providing feedback to indicate the strengths or the weaknesses of what has been posted. The comment can be substantive to the point that it rates nearly as high in importance as the original submission, or the comment can be a simple reinforcement of what was submitted previously. Ratings are brief and require minimal effort on the part of the visitor. As implemented in the Ensemble portal, ratings mean indicating a score of 1 to 5 stars, with the number of stars corresponding to the quality of the item in the opinion of the visitor. A rating that does not include a comment may not be of great value for an end user. However, due to their more formal nature, ratings provide the most reliable information for collaborative filtering and social navigation.

Tagging is the labeling of the item with key words and terms that allows the users to organize information. In the Ensemble project, we have provided both controlled vocabulary and free form (community) tagging. The advantage of the controlled vocabulary, here specifically the computing ontology [4], is that the same word or

phrase will be used consistently, aiding in effective search and classification. The advantage of community tagging is the ability of the user to express what they believe are the relevant features of the item without having to compromise their expression when the controlled vocabulary does not match their need very well. Both are valuable and we will explore the usefulness of both as the project progresses.

In the social navigation component of Ensemble, the CRT feedback is used in two ways. First, all of them indicate user interest in various items. Second, they allow the system to identify items, which are similar from the user behavior prospect. This data is used to generate links to similar resources.

4.3 Fragment Selection: Superimposed Information in Ensemble

One of the main goals of Ensemble is to enhance the application of stored information by allowing the users to use content both below and above the stored level. Ensemble provides support for *superimposed information* - the use of fine-grained fragments from existing data sources in new contexts. The main concept is that potentially fine-grained selections from an existing document or digital object can be referenced (using an appropriate addressing scheme) and then used in a variety of ways. We and our colleagues have built a variety of tools that allow users to select portions of MS Office, XML, PDF, and HTML documents and structure, annotate, and elaborate them in a scratchpad tool [5], for example, highlight and label appropriate selections from an electronic copy of a textbook with the appropriate learning objective from a university database course [1]. Our data shows that superimposed information is useful for end users and information access tools (i.e., to improve document [10]).

Ensemble implements support for superimposed information using Fedora. The implementation allows a user to easily create what we call subdocuments – the selected portion of an existing digital object – as a first-class object in a digital library. An author can select a portion of a web resource, annotate it graphically in situ, and then save the resulting annotated excerpt as an object in a digital library, along with metadata that specifies the address of the entire original resource. For each such object, we also create a view consisting of a web page that displays the annotated excerpt along with a link to the original resource. Our work enables all user actions against subdocuments to be logged. More than that, Ensemble can track the creation of subdocuments as well as their use in more complex, superimposed digital objects, including Waldens' Paths, described in the next section. The information about fragment extraction, annotation, and reuse provides valuable information for social navigation mechanisms of Ensemble.

4.4 Feedback by Composing: Waldens' Paths

In addition to the ability to extract information, Ensemble allows the user to compose information active *above* the storage level. This ability is supported by a guided path mechanism known as Walden's Paths [11]. A guided path is an organized and annotated collection of resources composed by an end user. Superficially a path has many common attributes with a resource list since it is essentially a collection of links with annotations like some resource lists. However, the path provides a collection of resources external to the path but still part of the path, unlike resource lists, the *in situ*

nature of its annotation makes it possible for an author to form a narrative structure while still containing the resources of interest.

This system provides users with authoring and viewing interfaces. The authoring interface allows users to compile their resources and add annotations in order to construct a narrative structure while maintaining the original form of the individual resources. The resources and their annotations create a linear structure to better facilitate the path creator's constructed narrative. After the user has completed authoring a path, they may choose to make it publicly available to others or may choose to provide the path's URL directly to the other users. The viewing interface allows users to view paths. When viewing a path, the user sees the original resource inside of the interface as well as an external hyperlink containing the original resource's URL. If the user navigates off of the guided path the navigational interface changes to notify the user that he has ventured off the guided path. At any point, if the user wants to return to the guided path from his self-guided exploration, he may click the "back to path" button to return to the point in the path where he last left off.

In practice, a path acts as a communication medium through which a creator can interact with a user. A creator can be a teacher providing education materials to a class or to other teachers, or a student creating a path to demonstrate their understanding of a topic to the teacher or other students—indeed in our experiences with Walden's Paths we have observed all of these cases. Outside of the formal educational setting, a path author can be any person who is interested in providing a constructed narrative around a set of web resources. In essence, Walden's Paths becomes the intermediary through which previously disconnected resources are contextualized and communicated. Being important in DL context for a range of reasons, Walden's Paths serve as an invaluable source of information for social navigation. The creator's decision to include a resource into a guided path indicates user and group interest in this resource, which can be used for social annotation. The co-location of items in various paths and their annotation provide a reliable way to estimate resource similarity from a perspective of a specific group. The discovery of this similarity is critical to generating links to recommended items.

5 The Mechanisms for Social Navigation

To account for the distributed nature of Ensemble, the portal uses a service-based personalization approach. All kinds of social navigation are provided by the personalization engine PERSEUS, which is independent from the portal. As a result, social navigation to portal resources can be provided not only inside the main portal, but also within its various components and tools (such as Walden's Paths) and even on personal sites of users who apply portal resources for their needs.

The PERSEUS engine is able to offer social navigation support for any list of links to resources inside or outside the portal. Inside the portal, these lists of links can be found in the terminal nodes (leaves) of the portal's traditional hierarchical organization. In this case, this list of links corresponds to a group of similar resources collected and classified by the portal. Outside the portal, this list of links can appear on a Web site of the instructor as resources assembled to support a specific lecture.

To extend this list of resources with social navigation support, the list should be included in an *embedlet*, a specific fragment of HTML, which is responsible for both

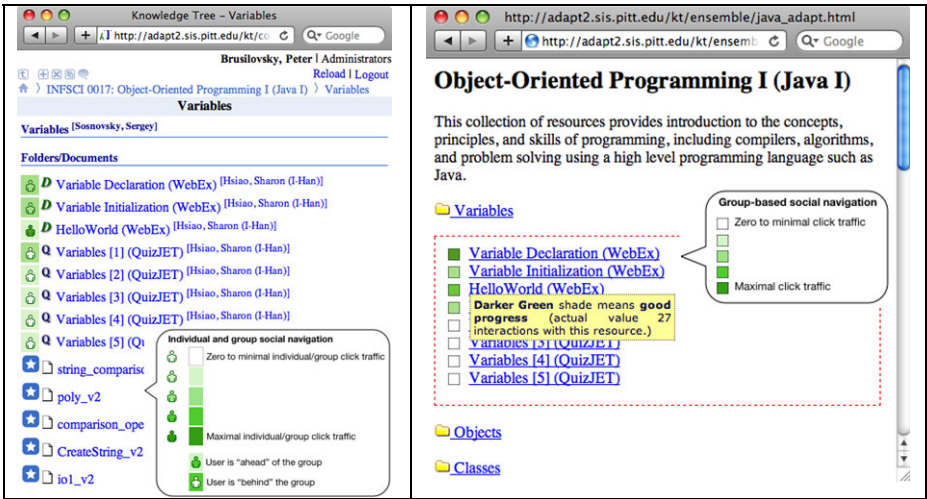


Fig. 2. Social navigation support on a portal (left) and a static web page (right)

tracking user browsing and generating social guidance. When a Web browser renders this embedlet, the list of resources along with the information about the current user and her group is sent to PERSEUS. PERSEUS contacts the user modeling server and produces two kinds of social navigation support for this list: social visual cues and personalized recommendations. *Social visual cues* annotate links in the original list showing their relevance to the current community. For example, Fig. 2 shows cues that indicate resource popularity by color intensity (the more intensive the color is, the more popular is the resource in a group). *Personalized recommendation* extends the original list with other resources, which are considered similar to the resources in this list from the perspective of the current group. To distinguish the original and the recommended resources, the links to recommender resources are annotated by the star icon (Fig. 2).

6 Summary

This paper argued for the need to extend digital libraries with social navigation features and demonstrated how social navigation can be implemented in the context of a large distributed educational DL, the Ensemble computing portal. We demonstrated a range of user interactions, which can be tracked to form the collective wisdom and presented an architecture, which is able to apply this wisdom to guide future users of Ensemble. The social navigation approach is already used in some components of the portal. The evaluation shows that even a single-server implementation can support much larger volume of users than we have now.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant Numbers 0534762, DUE-0840713, 0840719, 0840721, 0840668, 0840597, 0840715, 0511050, 0836940 and 0937863.

References

1. Archer, D.W., Delcambre, L.M.L., Corubolo, F., Cassel, L.N., Price, S., Murthy, U., Maier, D., Fox, E.A., Murthy, S., McCall, J., Kuchibhotla, K., Suryavanshi, R.: Superimposed information architecture for digital libraries. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 88–99. Springer, Heidelberg (2008)
2. Brusilovsky, P.: Social Information Access: The Other Side of the Social Web. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M., et al. (eds.) SOFSEM 2008. LNCS, vol. 4910, pp. 5–22. Springer, Heidelberg (2008)
3. Brusilovsky, P., Chavan, G., Farzan, R.: Social adaptive navigation support for open corpus electronic textbooks. In: De Bra, P.M.E., Nejd, W. (eds.) AH 2004. LNCS, vol. 3137, pp. 24–33. Springer, Heidelberg (2004)
4. Cassel, L.N., Davies, G., Fone, W., Hacquebard, A., Impagliazzo, J., LeBlanc, R., Little, J.C., McGettrick, A., Pedrona, M.: The computing ontology: application in education. In: Working Group Reports of Innovation and Technology in CS Education, pp. 171–183 (2007)
5. Delcambre, L.M.L., Maier, D., Bowers, S., Weaver, M., Longxing, D., Gorman, P., Ash, J., Lavelle, M., Lyman, J.: Bundles in Captivity: An Application of Superimposed Information. In: Proc. of 17th International Conference on Data Engineering, pp. 88–99. Springer, Rome (2001)
6. Dieberger, A., Dourish, P., Höök, K., Resnick, P., Wexelblat, A.: Social navigation: Techniques for building more usable systems. *Interactions* 7(6), 36–45 (2000)
7. Dieberger, A., Guzdial, M.: CoWeb - experiences with collaborative Web spaces. In: Lueg, C., Fisher, D. (eds.) From Usenet to CoWebs: Interacting with Social Information Spaces, pp. 155–166. Springer, New York (2003)
8. Farzan, R., Brusilovsky, P.: AnnotatEd: A social navigation and annotation service for web-based educational resources. *New Review in Hypermedia and Multimedia* 14(1), 3–32 (2008)
9. Farzan, R., Brusilovsky, P.: Where did the Researchers Go? Supporting Social Navigation at a Large Academic Conference. In: Proc. of The 19th ACM Conference on Hypertext & Hypermedia, pp. 203–211 (2008)
10. Price, S., Nielsen, M., Delcambre, L., Vedsted, P., Steinhauer, J.: Using semantic components to search for domain-specific documents: An evaluation from the system perspective and the user perspective. *Information Systems* 34(8), 724–752 (2009)
11. Shipman, F., Furuta, R., Brenner, D., Chung, C., Hsieh, H.: Guided Paths through Web-Based Collections: Design, Experiences, and Adaptations. *Journal of the American Society for Information Science* 51(3), 260–272 (2000)
12. Wolpers, M., Memmel, M., Giretti, A.: Metadata in Architecture Education - First Evaluation Results of the MACE System. In: Cress, U., Dimitrova, V., Specht, M. (eds.) Proc. of 4th European Conference on Technology Enhanced Learning (ECTEL 2009), Nice, France, pp. 112–126. Springer, Heidelberg (2009)

Automating Logical Preservation for Small Institutions with Hoppla

Stephan Strodl, Petar Petrov, Michael Greifeneder, and Andreas Rauber

Vienna University of Technology, Vienna, Austria
{strodl,petrov,greifeneder,rauber}@ifs.tuwien.ac.at

Abstract. Preserving digital information over the long term becomes increasingly important for a large number of institutions. The required expertise and limited tool support discourage especially small institutions from operating archives with digital preservation capabilities. Hoppla is an archiving solution that combines back-up and fully automated migration services for data collections in environments with limited expertise and resources for digital preservation. The system allows user-friendly handling of services and outsources digital preservation expertise. This paper presents the automated logical preservation process of the Hoppla archiving system in detail. It describes the recommendation process for appropriate preservation strategies via a web update service. A set of two real world case studies were conducted based on a first rules set focused on common office documents. The promising results sustain the novel approach of automating logical preservation by outsourcing expertise.

1 Introduction

Digital information forms essential assets in the long run for an increasing number of small institutions. Not only legal obligations mandate archiving of digital objects, but the loss of data can lead to serious business problems, e.g. data containing intellectual property, know-how, expertise or business data.

The common practice is regular backup on storage devices be they external hard disks or DVDs. Most users do not know the exact specification or format of their digital objects. Neither are they aware of changes in technological environment and therefore which formats could still be rendered in 5, 10 or even 15 years. Although the bitstream preservation problem is not entirely solved, there exist many years of practical experience in the industry, with data being constantly migrated to current storage media types, and duplicate copies held to preserve bitstreams over years.

A much more pressing problem is logical preservation. The rendering of a bitstream depends on the environment of hardware platforms, operating systems, software applications and data formats. Even small changes in this environment can cause problems in opening an object. Digital preservation is mainly driven by memory institutions like libraries, museums and archives, which have a focus on preserving scientific and cultural heritage, and dedicated resources available

to care for their digital assets. Enterprises whose core business is not data curation are going to have an increased demand for knowledge and expertise in logical preservation solutions to keep their data accessible. Long-term preservation tools and services are developed for professional environments to be used by highly qualified employees in this area. In order to operate in more distant domains, automated systems and convenient ways to outsource digital preservation expertise are required.

The Hoppla Archiving System¹ [9] combines back-up and fully automated migration services for data collections in small office environments. The system allows user-friendly handling of services and outsources digital preservation expertise. Hoppla uses its migration capabilities to continuously migrate digital objects which are in danger of being obsolete and unaccessible in the near future to more stable formats. The knowledge how and under which circumstances to migrate a certain object is provided by experts. This information is distributed to every Hoppla client upon request via a central web service.

The major challenges of an automated archiving system are the decision-making ability and the error tolerance of the software system. In particular the migration process is highly error-prone and needs special automated error handling mechanism due to the limited competence of users of troubleshooting and decision making. A further challenge is the variety of formats in the collections of small intuitions. The archiving system need to provide migration pathways for a large number of formats that are at risk of becoming obsolete. This paper presents the automated logical preservation process of the Hoppla system in detail.

The remainder of this paper is organised as follows: Section 2 provides pointers to related initiatives and gives an overview of work previously done in this area. Section 3 describes the automated logical preservation process of the Hoppla system, followed by the results from a first set of case studies in Section 4. An outlook on future work and a final conclusion is presented in Section 5.

2 Related Work

The concept and the design of the Hoppla system is presented in [9]. A number of research initiatives have emerged in the last decade in the field of digital preservation, primarily memory institutions focusing on professional environments. The raising awareness for small institutions and SOHOs increases demand of practical solutions for users with less experience [2].

Existing open source digital repositories, such as Fedora Commons² and DSpace³, are developed for large scale collections in professional archiving. These repositories provide a huge function range, but require considerable knowledge for configuration and usage. The overhead of function and configuration make

¹ <http://www.ifs.tuwien.ac.at/dp/hoppla>

² <http://www.fedora-commons.org>

³ <http://www.dspace.org>

these systems unsuitable for institutions with limited knowledge in data management. The innate support of these systems for logical preservation is limited. Considerable effort of integration and development would be necessary to provide long term preservation functionality for a collection. Another repository such as the e-Depot [8], developed by KB and IBM focus on electronic publications and is also developed for use in professional settings.

The CRIB project [3] has developed a Service Oriented Architecture implementing migration support. The digital objects are transferred to a server infrastructure and migrated objects are returned. The actual migrations of the objects are executed on the server side. CRIB is integrated into the RODA repository⁴. Transferring complete data collection to an external web service is potentially inefficient for small institutions and raises privacy issues.

The Panic Project [5] developed a framework to dynamically discover preservation strategies with similarities to the Hoppla system. Panic uses semantic web technologies to make preservation software modules available as Web services. The system is designed for large-scale repositories that implement the required services invoker. Panic uses external web services with actual data similar to the CRIB project. The Hoppla web service on the other hand provides only preservation recommendations to the client system. The migrations of the data are executed on the client side without transferring private data via the internet. Moreover the Hoppla system includes the bit preservation of the data.

The PreScan system [7] automatically extracts embedded metadata from digital objects. The system scans objects on a hard disc and manages their metadata in an external repository that supports Semantic Web technologies. The metadata could be used to implement digital preservation support.

A range of projects are developing tools and components to support digital preservation. Format registries that can be used to determine required preservation actions are developed by UK National Archive's PRONOM project [10] and the Global Digital Format Registry (GDFR) [4]. The format identifier tool Droid⁵, based on the Pronom registry, is used in the Hoppla system to determine objects format. In order to obtain additional metadata about objects, a number tools and services are developed. For example, JHove⁶, developed by JSTOR and the Harvard University Library is used within Hoppla.

Research on technical preservation issues is focused on two dominant strategies, namely migration and emulation. The Council of Library and Information Resources (CLIR) presented different kinds of risks for a migration project [6]. Migration requires the repeated conversion of a digital object into more stable or current file format. Migration is a modification of the data and always incurs the risk of losing essential characteristics of the object [6]. Still the number of tools as well as the ease of applying migration makes it a very promising candidate for archiving in small institutions. Emulation, the second important preservation strategy aims at providing programs that mimic a certain environment. The

⁴ <http://roda.di.uminho.pt>

⁵ <http://droid.sourceforge.net>

⁶ <http://hul.harvard.edu/jhove>

major disadvantage is that the emulator itself is a piece of software and has to be preserved over time. In order to keep the archiving system simple and easy to apply, we are currently not considering emulation as a preservation strategy for Hoppla.

3 The Hoppla Logical Preservation Process

The workflow and the design of the Hoppla system is presented in [9]. In this paper the automated logical preservation process of the Hoppla archiving system will be presented in detail. The basic idea is to provide preservation know-how to client users with limited expertise. Therefore the client sends information about the collection, the available migration tools and the user's requirements to a central update service (Figure 1). Based on the information appropriate preservation rules and tools for objects that are at risk of becoming obsolete are selected by the update service and send to the client. Migrations of the objects are executed according the provided rules on the client side. The process is described in detail in the following sections.

3.1 Collection, User and System Profiling

The first step in the process is to create a collection profile on the client side. The profile describes the technical characteristics of the collection to be preserved. The description includes the objects formats as well as more detailed description, for example the encoding of videos, size of the object, transparent layers of images, etc. This information is essential to select suitable migration strategies. The profile consists of aggregated metadata extracted from digital objects in the collection. The metadata are created within Hoppla by using format identification tools (e.g. DROID) and characterisation tools (e.g. JHOVE).

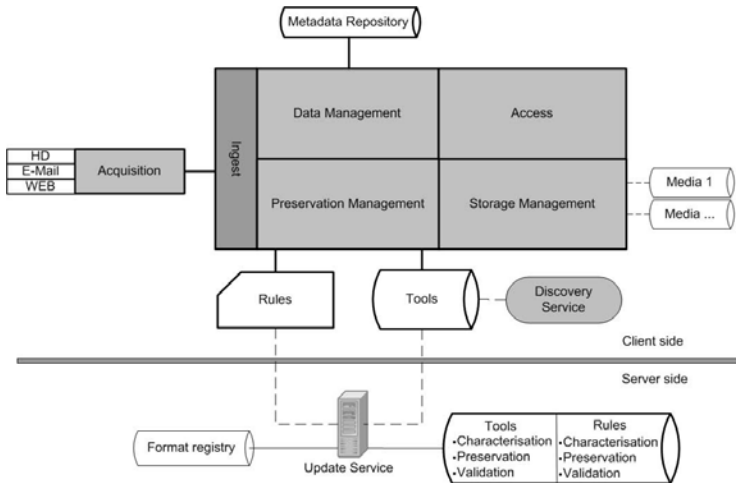


Fig. 1. Architecture of the Hoppla System

User's requirements on the digital data differ significant between user groups. These requirements and particularly future usage scenarios of the digital objects have substantial influence on the selection of preservation rules. For example a professional photographer has other future usage scenarios for images (enlarging, cutting, editing) than other users and therefore need other migration rules. In order to capture the preferences in a user-friendly way Hoppla has introduced a logical preservation level rating for four types of objects (text, audio, video and images). The rating ranges from zero (no preservation actions) to one (all available actions). According the rating preservation rules are selected. For a low rating only essential preservation actions for objects that are at imminent risk of becoming obsolete are performed. For a medium rating migrations are done for formats that are no longer in wide-spread use, outmoded (e.g. old versions of application) or the migration of proprietary format to open source formats. High ratings initiate additional multiple migration pathways for object formats to support different future usage scenarios e.g. migration of Microsoft Word objects to PDF/A, OpenOffice Format(ODT) and plain text format (preserving the layout, the editability and allowing computer-assisted processing). The selection of the preservation strategies is presented in detail in Section 3.3.

In addition to collection characteristics and user preferences the technical environment on the client side effects the selection of preservation rules. It is described in the system profile. The profile consists of technical characteristics (such as operating system, free storage size, etc.) and installed migration tools on the client system. A main challenge of an automated archiving system is the provision of useful migration services for large number of object formats. Hoppla supports two kinds of migration service, portable and external. Portable migration services are provided by the update service. The services are downloaded and dynamical integrated into the Hoppla client application. This mechanism works excellent for limited number of migration tools, for example tools implemented in Java. In order to provide migration support for a larger variety of formats Hoppla uses migration services installed on the client system (so called external services). A service to discover installed tools is described in the next Section 3.2.

3.2 Tool Discovery

In order to provide migration paths for a wide range of formats Hoppla uses migration services installed on the client host system. A discovery service helps to identify tools available on Linux or Windows systems. The current version of the discovery tool only supports tool that can be executed via the command line. The migration services to discover are defined in a XML configuration file, which can easily be updated through the web service. Examples for migration services are ImageMagick⁷, OpenOffice⁸, Mencoder⁹.

⁷ <http://www.imagemagick.org>

⁸ <http://www.openoffice.org>

⁹ <http://www.mplayerhq.hu>

For Linux the system executes the command and parses the outcome to check whether the tool is installed on the system and which version of the tool. Typical installation paths (e.g. /usr/bin/) can be specified to allow a more precise search. On Windows machines the system registry is searched for migration tools and their installation paths. The result of the discovery services is a list of installed migration tools including their installation path on the system. These tools can be used by the Hoppla system for migrations.

3.3 Selection of Preservation Strategies

One of the most challenging tasks in the preservation process is the selection of appropriate preservation strategies for a given collection. A profound and solid evaluation of different preservation alternatives is a time-consuming task and requires input by different domain experts (e.g. technic, curation, etc.). Work on preservation planning has been done for example with the planning tool Plato [1]. It was developed to support the evaluation of different potential preservation strategies against individual preservation requirements. The requirements are collected from the wide range of stakeholders and influence factors that have to be considered for a given setting. The Plato approach was designed to be used in professional settings by preservation expert. Due to the limited expertise and knowledge of the typical Hoppla user an expert preservation planning process (such as Plato) is not feasible and an alternative approach is required. The Hoppla Update Service implements a preservation recommendation service enabling outsourcing of the selection of appropriate preservation strategies. It provides best practice digital preservation rules for individual collections of Hoppla users.

Preservation Rules of the Update Service. The preservation rules of the Hoppla Update Service are created and administered by a group of preservation experts. A screenshot of the rule administration web site is shown in Figure 2. The rules in the system base on experience in professional settings and intensive testing and evaluation based on test corpus. Unlike in professional settings where preservation rules are created for a specific collection, the rules in the update service are applied to a large number of collections of a specific format. Migrations used in automated migration setting need to provide a high degree of robustness and error tolerance.

Extensive testing and evaluation of the preservation rules is required by preservation experts. For sound testing extensive data corpora of specific formats are required. The objects need to represent the different technical characteristics that a format can have. For example the test objects can vary in size, embedded objects, transparent layer in images, different codecs, etc. The different technical characteristics can have significant effects on the outcome of migrations. Advanced corpora should also contain objects that do not meet the format specification to test the error tolerance of the migration process.

The evaluation and selection of the preservation rules for the web update service is a very challenging task. General preservation requirements and evaluation criteria that apply to all Hoppla users (or even a user group) have to be found.



Fig. 2. Web Service Rule Management

Best practices, operating experience of professional archives and de-facto standards (e.g. PDF/A) support the decision making. Through the generalisation not all individual requirements of single users can fully met by the system. Nevertheless the Hoppla system provides best effort preservation rules for specific collections.

As the documentation of the whole process in an archive becomes more important for audit and certification initiatives the web service provides a preservation plan for each preservation rule documenting evaluation criteria, compared alternatives, test objects, potential losses and evaluation results.

A preservation rule of the Hoppla Update Service consist of

- **unique identifier**
- **label & description of the migration strategy**
- **migration tool & applied parameter** including description of tool (licence, operating system)
- **source format and constraints** are specifying in detail the technical characteristics objects covered by the rule need to have
- **target format extension** is required to name the resulting object according to the target format
- **preservation level** specifies the required preservation level (for text, audio, video and images) to apply this strategy
- **estimated duration and size change**, is used to forecast the duration of the migration process and calculate the additional storage usage
- **validity period** of the rule
- **documentation** is a set of documents containing a preservation plan for the preservation rule (evidence of traceability and accountability of the logical preservation in the Hoppla system)

The actual selection of preservation rules for a specific request is based on the collection, user and system profile. In a first implemented version the server selects all rules that are available for the formats specified in the collection profile. After that all rules are selected that conform to the preservation level. Finally, the required tools for the chosen rules are checked whether the tools are installed or available for download from the server, resulting in a list of rules that are sent to the client of Hoppla in XML format.

The recommendations provided by the web update service are best effort approaches for automated digital preservation. It can not substitute individual expert planning processes for professional digital preservation endeavours. However, it can provide a practical way for small institutions with limited in-house resources to preserve their digital collections.

3.4 Application of Preservation Strategies

The last step of the workflow is the execution of the recommended preservation strategies on the client side. All objects in the collection that are covered by a preservation rule are identified. The list of migrations is presented to the user. More advanced users can decide which migrations to perform and which to cancel. Furthermore, the duration and additional storage usage of the migrations are estimated based on the figures in the migration rules and the size of the objects in the collection.

The next step is the preparation of the migration services. Portable migration services that are required for the selected migrations are transferred to the Hoppla client. The update service provides a resource based service to download the services, the URLs of the required services are specified in the rules.

The objects in the archive are migrated according to the rules on the client side. Migration is a critical task, as migration services tend to be very error-prone. The large variety of characteristics of a single format that do not always conform to official format specifications (e.g. the use of undocumented features of a format) can cause malfunction of services. Special error handling mechanisms are implemented in the Hoppla software to deal with services that block, abort the migration process or produce unexpected errors. The migrations are executed in a temporary directory on the client system. Successfully migrated objects are added to the existing collection in the data management and metadata about the new objects are collected. In a final step the migrated objects are stored on the storage media of the system.

As privacy is an important aspect of outsourcing it needs to be mentioned that the actual data are not transferred and all actions on the data (migration, identification, metadata extraction) are executed on the client side. Moreover in the next version of Hoppla the level of detail and the containing information of the three profiles (user, collection and system) will be selectable by the user. This should provide the highest degree of transparency for information shared with the web update service.

4 Case Study

A series of two case studies was conducted with a special focus on the logical preservation capacities of Hoppla. In the first case study data of research projects were preserved, the data primarily consists of common office formats. In the second study a business e-mail account was preserved. A common set of rules was applied in both studies with the objective to provide well established and practicable preservation for common office formats (such as text, presentation and spreadsheets). The experiments were executed on workstation PC with a Windows XP operating system.

In our rule set we demonstrate the Hoppla capacity for multiple migration paths for a single format (in our rule set for Microsoft Word documents). The multiple paths should ensure the highest possible support for different future usage scenarios (for example PDF/A for printing and viewing, OpenOffice format for later editing and plain text for text retrieval operations).

Rule set

The following migration rules were applied to both collections:

- **DOC(X)2PDF.** Migration of Microsoft Word documents (objects with .doc and .docx extension) to Adobe PDF/A documents using the Java OpenDocument (JOD) Converter¹⁰ converter. The JOD converter uses an OpenOffice instance¹¹ to perform the document conversions. On both of our test systems OpenOffice 3.1.1 was installed.
- **DOC(X)2ODT.** Migration of Microsoft Word documents to OpenOffice document format using the JOD converter.
- **DOC(X)2TXT.** Migration of Microsoft Word documents to plain text using the JOD converter.
- **PPT(X)2PDF.** Migration of Microsoft PowerPoint documents to Adobe PDF/A documents using the JOD Converter
- **XLS(X)2ODS.** Migration of Microsoft Excel documents into OpenOffice Calc Files using the JOD Converter
- **FLV2MPG.** Migration of FLV video (Flash Videos) to MPG videos (compressed version using MP3 Audio codec and x264 (MPEG-4 Advanced Video Coding (AVC))video codec) using MEncoder¹²
- **WAV2FLAC.** Migration of WAVE documents to Flac using Flac 1.2¹³
- **PS2PDF.** Migration of PostScript documents to Adobe PDF documents

4.1 Sample Set Office Documents

The sample set contained the data of 5 projects (3 small one and 2 multi-year research projects). The size of the set was 2 GB of data and contained about

¹⁰ <http://artofsolving.com/opensource/jodconverter>

¹¹ <http://www.openoffice.org>

¹² <http://www.mplayerhq.hu>

¹³ <http://flac.sourceforge.net>

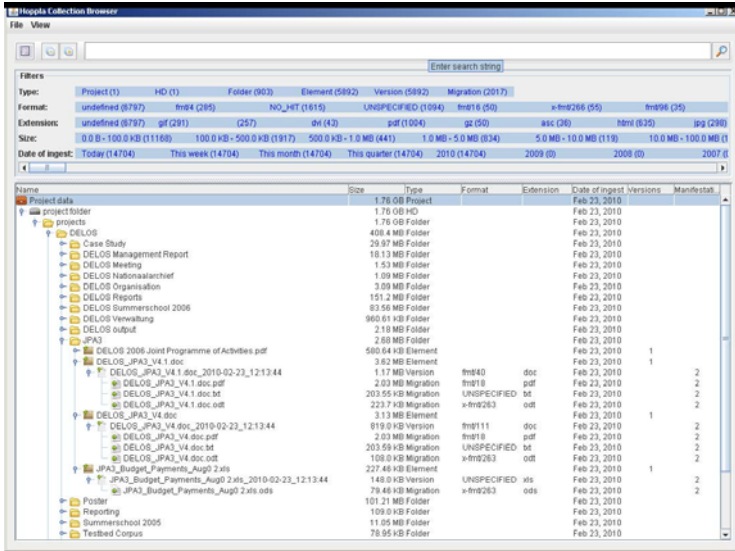


Fig. 3. Collection Browse Projects

6,000 objects in 1000 folders. The collection mainly consists of objects in office formats (e.g. about 524 doc (110MB), 250 xls (160MB), 1000 pdfs (500MB), 500 png (230MB)). Two external hard discs were used as storage media.

Hoppla recommends 2045 migrations and estimated the additional storage need at 660MB. The preservation process (including acquisition, migrations and storage) took about 1,5h. Figure 3 shows the Hoppla collection browser with the resulting collection presenting the multiple migration paths for Microsoft Word documents.

Hoppla successfully migrated 2017 objects, the resulting migrated objects had a size of 380MB. 28 migrations were reported as failure, in particular the migration of PostScript documents to Adobe PDF documents causes ten failures. The implementation of the ps2pdf service on Windows seems to be unstable and required re-evaluation of the rule and the service. The other failures were mainly caused by corrupt documents. The size estimation for the migrated objects was inaccurate and need also further adjustment. Some results in detail, the 524 objects in Microsoft Word format with size of 110MB resulted in 6MB txt documents, 165MB PDF/A documents and 60MB of ODT objects. 248 documents in XLS format (160MB) were migrated into 24MB of ODS format documents. Three documents in flv format with a size of 56MB were migrated to MPG videos according to the migration rules (78MB). The preserved collection on the storage media including the migrated objects and the metadata of the collection in xml format (7MB) had a size of 2,3GB.

4.2 Sample Set E-Mails

The second case study dealt with an e-mail repository containing about 2000 e-mails. About 600 e-mails had attachments, the most common format were Adobe PDF (268), Microsoft Word (145), JPG images (100). The e-mail including the attachments had a size of 271MB, the average size of the attachments was about 300KB. Hoppla stores the e-mail header and content as plain text files. Here, the Hoppla software need to overcome numerousness of different e-mail encodings. The attachments of the e-mails are stored in separate folders. Hoppla identifies the format of the attachment and requested preservation rules from web update service for obsolete formats.

In our case study the Hoppla software performed 493 successful migrations. Eight migrations failed because three documents were corrupt. The migrated object had a size of 73MB. The migration of all objects took about 20 minutes. A manual inspection of some random migration outcomes showed very positive results. The resulting PDF and ODT documents from migrations of Microsoft Word documents were completed and correct. Even the migration to plain text produced satisfactory output. Large parts of the containing text could be extracted.

5 Conclusion

In this paper we described the automation of logical preservation within the Hoppla system. Appropriate preservation strategies for specific collections are recommended by a central web update service. The selection is based on information about the collection, the user and the system which are provided by Hoppla clients. In order to provide migration paths for a wide range of formats Hoppla includes a discovery service to identify potential preservation services on the client side. The recommended migrations are performed on the client side and the migrated objects are managed by the Hoppla system and stored on external media.

Two realistic case studies produced good results and indicated the applicability of the approach. The first set of preservation rules provided migration paths for common office formats. The migration failures were mainly caused by corrupt objects. Only the ps2pdf migration service on windows seems to be unstable and not reliable.

This first version of Hoppla presents the fundamental progress, that will be further refined. The robustness of the migration strategies has to be investigated in more detail and the detection of corrupt objects should be improved. Moreover potential preservation strategies and tools for more formats have to be identified and integrated. A second version of the selection process for migration strategies will allow taking into account more information from the client and thus offering more accurate recommendations.

Acknowledgements

Part of this work was supported by the EU in the 6th FP, IST, through the Planets project, contract 033789 and the Research Studios Austria program of the Federal Ministry of Economy, Family and Youth of the Republic of Austria.

References

1. Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H.: Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* (2009)
2. Bradley, K.: Digital Preservation: the need for an open source digital archival and preservation system for small to medium sized collections (2008), <http://portal.unesco.org/ci/en/files/28067/12323631793BradleyPaper.pdf/BradleyPaper.pdf>
3. Ferreira, M., Baptista, A.A., Ramalho, J.C.: An intelligent decision support system for digital preservation. *International Journal on Digital Libraries* 6(4), 295–304 (2007)
4. Harvard University Library. Global digital format registry (GDFR), <http://hul.harvard.edu/gdfr>
5. Hunter, J., Choudhury, S.: PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. *International Journal on Digital Libraries: Special Issue on Complex DigitalObjects* 6(2), 174–183 (2006)
6. Lawrence, G.W., Kehoe, W.R., Reiger, O.Y., Walters, W.H., Anne, K.R.: Risk Management of Digital Information: A File Format Investigation. In: Council on Library and Information Resources (2002)
7. Marketakis, Y., Tzanakis, M., Tzitzikas, Y.: Prescan: towards automating the preservation of digital objects. In: MEDES 2009: Proceedings of the International Conference on Management of Emergent Digital EcoSystems, pp. 404–411. ACM, New York (2009)
8. Oltmans, E., van Diessen, R., van Wijngaarden, H.: Preservation functionality in a digital archive. In: JCDL 2004: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 279–286. ACM Press, New York (2004)
9. Strodl, S., Motlik, F., Stadler, K., Rauber, A.: Personal & SOHO archiving. In: Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries (JCDL 2008), Pittsburgh PA, USA, pp. 115–123. ACM, New York (2008)
10. The National Archives. Pronom - the technical registry, <http://www.nationalarchives.gov.uk/pronom>

Estimating Digitization Costs in Digital Libraries Using DiCoMo

Alejandro Bia¹, Rafael Muñoz², and Jaime Gómez²

¹ CIO/DEMI, Miguel Hernández University, Spain
abia@umh.es

² DLSI, University of Alicante, Spain
{rafael,jgomez}@dlsi.ua.es

Abstract. The estimate of digitization costs is a very difficult task. It is difficult to make exact predictions due to the great quantity of unknown factors. However, digitization projects need to have a precise idea of the economic costs and the times involved in the development of their contents. The common practice when we start digitizing a new collection is to set a schedule, and a firm commitment to fulfill it (both in terms of cost and deadlines), even before the actual digitization work starts. As it happens with software development projects, incorrect estimates produce delays and cause costs overdrafts.

Based on methods used in Software Engineering for software development cost prediction like COCOMO and Function Points, and using historical data gathered during five years at the Miguel de Cervantes Digital Library, during the digitization of more than 12.000 books, we have developed a method for time and cost estimates named DiCoMo (Digitization Costs Model) for digital content production in general. This method can be adapted to different production processes, like the production of digital XML or HTML texts using scanning and OCR, and undergoing human proofreading and error correction, or for the production of digital facsimiles (scanning without OCR). The accuracy of the estimates improve with time, since the algorithms can be optimized by making adjustments based on historical data gathered from previous tasks.

Keywords: Cost and time estimates, Digitization, Contents Production, DL Project management.

1 Introduction

Almost three decades after Barry Boehm presented the Constructive Cost Model (COCOMO) [3], the problem of accurately estimating software development costs is far from solved. In professional software development practice, just a few developers use software estimation methods other than expert judgment (which is basically an “expert’s guess”), and when they do, the results are usually far from satisfactory [11,9].

This work discusses some of the reasons why cost estimate methods like COCOMO fail in practice in software engineering applications, but may be accurate

for other tasks, like predicting digitization times and costs, provided we make the necessary modifications and customizations to the algorithm. By doing this, we have improved the accuracy of the estimates and widened its possible uses to other fields. Hence, we recommend the use of this type of algorithmic method for tasks other than software development, like DL contents production. Below we provide examples of production time and cost estimates obtained in this way at the Miguel de Cervantes Digital Library (MCDL)¹.

1.1 The Basic Digitization Cost Model

In the digitization cost model we propose, we use an equation similar to Intermediate COCOMO², but with some differences:

- **Size-Independent Overhead.** We added a new term called *SIO* (*Size-Independent Overhead*) that represents the fixed preparation time for the task, which is independent from its size. An example of this size-independent overhead is the time needed to adjust the parameters of an image scanner and OCR before starting a scanning session. This is a fixed time which does not depend on the number of pages to be scanned later.
- **The size is known beforehand.** One of the reasons why COCOMO often fails in estimating software costs is because its calculations are based on an estimated size of the code to be built (KLOC²), which is highly uncertain at the initial stages of the project. When applying a similar method to estimate digitization costs, the first thing we realize is that we don't have to guess the size of the work because we can easily know it, or can accurately estimate it. The size of the documents to digitize is measured as the number of pages P and can be measured (or calculated with reasonable accuracy) beforehand.
- **Time is cost.** There is one similarity with software development projects: since most of the cost in digitization is human labour, which in the long run overweighs the cost of hardware and software, time estimates of a digitization task can be directly converted to cost estimates, using some money-per-hour factor.

Given the number of pages P , we can directly calculate the time in hours T , with the Basic-DiCoMo formula:

$$T = a \cdot P^b + SIO \quad (1)$$

For a graphic example of this DiCoMo approach, see figure¹, where an estimation curve (thick line) approaches real data spots (black squares) that represent time measures of real digitized documents. The thin straight line represents a linear approach to the spots ($a \cdot P$), which is not the best approach, while the curve ($a \cdot P^b$) fits more accurately, although not perfectly. The values for a and b are obtained by adjusting the curve to best approach the cloud of points, using historical data. The value of the fixed term SIO is the point at which the curve

¹ <http://www.cervantesvirtual.com/>

² Kilo Lines Of Code

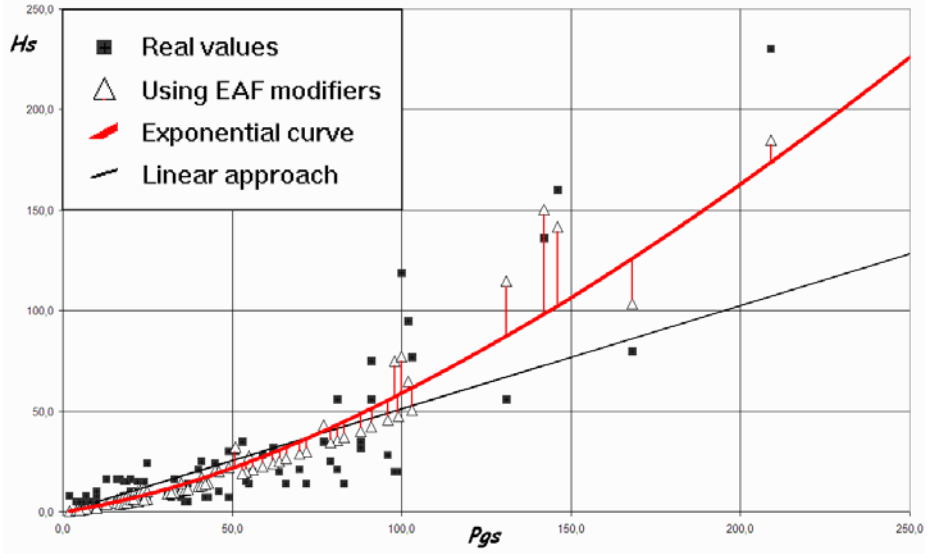


Fig. 1. Document digitization times (hours vs. pages)

crosses the Y axis, or the time needed for the impossible case of a task of size 0, and for our purposes represents the preparation time mentioned before. For certain tasks (e.g. big tasks), this time may be negligible and hence ignored.

For example, the following equation, based on early experience at the MCDL, gives us the estimated number of hours to process a text given the number of pages:

$$T = 0.069 \cdot P^{1.465} + 0.6 \tag{2}$$

Using this formula, a standard-complexity book of 100 pages will take about 59 hours of scanning, correction and XML markup altogether.

1.2 The Importance of Historical Cost-Data

Inside most organizations, the estimation of production costs is usually based on past experiences. Historical data are used to identify the cost factors and to determine their relative importance within the organization [8]. Historical data will be used first to adjust the basic estimation algorithm (the exponential curve), and later to adjust the detected impact factors to be used as modifiers to obtain more accurate results. This is the reason why it is so important to systematically collect and store time and feature data from projects, and to take note of the perceived factors that affect the times as well as the amount of this impact.

1.3 Adjusted DiCoMo

The simple approach used in equation (1) doesn't take into account the fact that different literary works have different degrees of difficulty owed to several factors

(discussed below), which will affect production times. We have detected the most important of these factors, and assigned weights to them to be able to use them as feature-modifiers. We added an Effort Adjusting Factor (*EAF*) to the Adjusted DiCoMo equation, equivalent to the one used in Intermediate-COCOMO, but based in this case on specific digitization features. The *EAF* is calculated as the multiplication of relevant feature-modifiers chosen from a table (see for instance table II). The modifiers shown in the table were obtained from historical data collected at the MCDL. The value of these modifiers is 1.00 in the normal case, then having no impact on the overall *EAF* factor, or values slightly above or below 1 in the other cases, contributing to raise or lower the unadjusted estimate, producing the desired “adjust” effect (see vertical lines from the exponential curve to the white filled triangles in figure II) .

$$T = a \cdot P^b \cdot EAF + SIO \quad (3)$$

where: $EAF = \prod modifier_i$

Table 1. DiCoMo: Complexity modifiers used to calculate the *EAF*

Modifier	Low	Normal	High
encoder experience and skills	1.30	1.00	0.70
familiarity with task	1.20	1.00	0.80
familiarity with computer tools	1.20	1.00	0.80
foreign or ancient languages present	—	1.00	1.25
stained or old paper	—	1.00	1.15
old font faces	—	1.00	1.15
special care required (ancient books)	0.80	1.00	1.20
high quality demands	0.80	1.00	1.20
inadequate technology used	0.80	1.00	1.20

1.4 Factors That Affect Digitization Costs

There are several factors that affect the cost of production of digital objects. Both these factors and their effect on costs are difficult to determine and have to be carefully studied. They are detected by experience, as features which are found to affect the time required to complete a task either positively or negatively. Once a factor of this type is detected, we have to measure its impact, as a percentage relative to the “normal-case” time. The best way to do this is by gathering time records of digitization tasks, and record also their particular features and their weight (e.g. low, normal, high). With enough records of this type, algorithm optimization techniques can be applied to infer the range of impact of a given feature as a +/- percent. For instance, we detected that the literary style of a text affected its digitization time, due to harder or simpler markup requirements. We started recording this feature, indicating whether a text was mainly, from best to worst case: prose, verse, drama written in prose or drama written in verse. We stored records of this together with the times required to complete the

digitization task. After gathering a good number of records, we used optimization techniques to get the optimum value range for this new modifier, which turned out to be +/- 7,6%. So in the case of drama written in verse (hardest markup case), we will have to add 7,6% more time to the estimate.

Among the factors detected, we can highlight the individual skills and experience of the persons assigned to the project, as well as their familiarity with the specific characteristics of the work to be digitized, the familiarity with the computer tools to be used, the complexity of the task, size, quality requirements, technology used, etc. Also important are some features of the document that affect digitization times like: the presence of foreign or ancient languages, stained/yellowish paper, old/irregular font faces, high quality demands, inadequate technology used, special care required for old books, etc.

The Adjusted DiCoMo equation (3), customized with historical data from previous projects (4), and using the EAF factor, now gives us better estimates of the time needed to digitize a text given the number of pages:

$$T = 0.081 \cdot P^{1.462} \cdot EAF + 0.1 \quad (4)$$

For instance, a book of 100 pages with stained/old paper (+15%) and foreign or ancient languages present (+25%), will take approximately 98 hours to complete, compared to the 59 hours estimated using the basic equation without modifiers: $T = 0.081 \cdot 100^{1.462} \cdot 1.15 \cdot 1.25 + 0.1 = 97.85h$

Figure 1 shows the Basic DiCoMo exponential curve (thick line), that approaches the black square data spots that represent measures of real digitized documents. The EAF adjusted results are shown as white filled triangles which in most cases approach more closely the real values. In a very few cases, however, the EAF results are worse than the basic curve.

The time assigned affects mainly the quality of the product obtained which is notably reduced when the times assigned are unreasonably short, forcing the technicians to work under excessive pressure. This is particularly true for the correction and editing process, where text output from OCR has to be carefully proofread and corrected. This is a delicate craft that takes time and cannot be done under excessive pressure. When not properly done, further revisions and corrections are needed, with a very negative impact on costs. Next, each one of these factors is described in detail.

1.5 Size of the Material to Publish

Digitization projects, compared to software development projects, have the advantage that we can know quite precisely beforehand the size of the work to be done (namely the number of pages or words to digitize).³

³ In software development projects, the number of lines of code is not known at the beginning of a project. This is the main drawback of the original COCOMO method, which was modified and renamed as COCOMO-II [4][5][6] to sort this problem. Other methods, like Function Points [1], Use Case Points [2][10] and Object Points [12], which are based on functionality aspects instead of lines-of-code, do not have this problem.

There are various ways to measure the size of the material to digitize. The first and easiest way to determine the raw size of a text to be digitized is to count the pages. This is the most common method, and is generally sufficient for accurate estimation purposes. A disadvantage is that pages are not equally dense for all books. We can have an approach to the density by counting the words that fit in a standard page, or the words that fit in a fixed size window, and then assuming that the rest of the pages are similar in this respect. To count individual words would be more accurate (we verified this by experience), but it is not a practical approach: the improvement in accuracy does not justify the effort. However, after the OCR process takes place, we will obtain a text file, with errors, but nevertheless a text file where we can automatically count the number of words or get the size in bytes. This is a good measure of the size of the proofread and correction work that follows, and may serve to adjust the initial estimates for higher accuracy.

1.6 Effort Adjusting Modifiers

There are many complexity factors that affect every stage of the digitization process (scanning, proofreading or correction, and markup). In the case of the correction stage, which we consider the most critical one, there are various factors to be taken into consideration:

- the type of text: prose, verse, drama (both written in prose, and in verse), dictionary.
- footnotes, if the number is too high
- quotations in foreign or classical languages (if too many)
- the complexity of the author style and vocabulary
- the quality of the OCR output (few or lots of errors)
- the legibility of the original (paper copy from which the digital version id produced).

Concerning markup, complexity varies according to the number and difficulty of the tags to be added. Drama, for instance, with the need of a *castlist*, *speaker* and *speeches*, require an additional amount of tagging compared to prose.

Verse with split lines is another good example of extra complexity, since special care needs to be taken to assign attribute values which indicate which part of the split line of verse is which (initial, middle, or final).

In the case of the production of digital facsimiles from manuscripts, a case of particular complexity is when we have to work on rare and valuable originals that have to be handled with special care (wearing rubber gloves for instance) and using a digital photographic camera instead of a flat bed scanner. On the contrary, digitizing unbounded pages using a flat bed scanner with automatic page feeder would be the easiest case.

For each of the critical features mentioned, three possible modifier values were set, to be used when the feature appear as high, normal, or low (e.g. the values could be 1.10, 1.00 and 0.90 in each case). For a given task, all the modifier values that apply to the case, and that are different than normal (1.00), are multiplied to obtain the *EAF* factor.

Individual skills of the technicians: In the programmer's world, individual productivity has been measured extensively. Harold Sackman et al. carried out an experiment in 1968 [13]. They made evident that performance differences registered in individual programmers were much bigger than those attributed to the effect of the working environment. The difference between the best and the worst performance was very high, being the experience a decisive factor. In a later experiment, Sackman observed a variation in the productivity of as much as 16 to 1. DeMarco and Lister also discussed the effects of a well integrated group to enhance productivity in their book *Peopleware* [7], that deals with the human component in software projects.

In digitization, the results that we have measured comparing correctors' performances show remarkable differences in productivity, depending on their individual skills and experience (sometimes a 3 to 1 ratio). Variations in productivity of this magnitude are significant for cost estimates, making necessary to express this in the calculations by means of a modifier.

Special quality requirements: In the case of digital text production, producing a modernized digital edition from an ancient text takes additional time and effort compared to processing a modern text, since modernization is a complex task that involves difficult decisions.

Using Madison markup for the transcription of a manuscript is another example of additional requested complexity. So is the case of making highly legible digital facsimiles from ancient manuscripts, where special care and fine-tuning of the scanning equipment may be required, as well as graphic postprocessing.

Technological level of the environment: This is a relevant issue when using different technologies or migrating from old to new production tools. When the environment is stable and well known, and the estimate equations are well adjusted for it, there is no need to care about this issue. Changes in technology, however, will surely require modifications to the equations, and may make historical time and cost data obsolete for future estimate adjusting purposes.

1.7 Procedure to Estimate Costs Using DiCoMo

1. Establish the production process to follow (production workflow). There may be different production workflows for different purposes (e.g. facsimile images are only scanned, while text undergoes scanning, OCR, proofreading and markup).
2. Identify all the objects (books, images, etc) to be digitized and their associated tasks (Work Breakdown Structure).
3. Measure or estimate the size of each object to be digitized.
4. Establish the production steps to be followed by assigning the right workflow.
5. Specify the effort adjusting factors for each object.
6. Calculate the time each unit will take (use the adequate equation with the corresponding complexity factors).
7. Calculate the total development time for the project as the sum of individual times.

8. Optionally, compare the estimate with another, perhaps a top-down one like the *DELFI* technique or *expert-judgment*, identifying and correcting the differences in the estimate if necessary.

1.8 The Most General Formula

In previous examples, we have used a single formula to estimate the whole digitization task, which is simpler, but better results can be obtained by using specific formulas with their own adjusts for each step in the digitization process (e.g., scanning, proofreading, markup). So in this case we consider each production step as a functional unit, to which a specific estimation equation is applied. The global estimate T turns out to be the sum of all the specific step estimates.

$$T = \sum_s (a \cdot P^b \cdot \prod eaf_i + SIO) \quad (5)$$

2 Implementation

The DiCoMo method was implemented into the digital library's workflow-and-document-handling system, a software application that controls the whole production process of all the types of digital resources produced by the DL. It provides useful management information for estimating costs and times of digitization projects. It estimates times of cataloging, scanning, correction and

Estimación de tiempos - DIGITALIZACION

Id: jcarlos | Id.Creador reg.: mamen
 Nombre: Juan Carlos García Candela

Núm.reg.: 7978
 Título: Visión de Andalucía

Tipo obra: [] | Autor/es: Basave Fernández del Valle, Agustín
 Tipo material: Obras modernas desde 1831
 Soporte: Papel

Número de páginas: 71 | Dispositivo utilizado: Escáner con ali...
 Dificultad tratamiento texto: Media | Dificultad tratamiento: Baja
 Dif. tratamiento imagen: Baja | Ajuste de tiempo: 0

Observaciones e incidencias: []

T. estimado: 284 4.44
 T. almacenado: 284 4.44

Insertar | Modificar | Generar ficha | Cancelar

Fig. 2. Estimate of scanning costs

Estimación de tiempos - DIGITALIZACIÓN

Id: Nombre:

Tiempo base:

	Baja	Media	Alta	min./pág.
Dificultad tratamiento texto	<input type="text" value="0.0"/>	<input type="text" value="3.0"/>	<input type="text" value="7.0"/>	<input type="text"/>
Dificultad tratamiento imagen	<input type="text" value="0.0"/>	<input type="text" value="6.0"/>	<input type="text" value="10.0"/>	<input type="text"/>
CD	<input type="text" value="1.0"/>	<input type="text" value="4.0"/>	<input type="text" value="7.0"/>	<input type="text"/>
Escáner sin alimentador	<input type="text" value="1.5"/>	<input type="text" value="7.0"/>	<input type="text" value="10.0"/>	<input type="text"/>
Escáner con alimentador	<input type="text" value="1.0"/>	<input type="text" value="4.0"/>	<input type="text" value="7.0"/>	<input type="text"/>
Escáner cenital	<input type="text" value="5.0"/>	<input type="text" value="10.0"/>	<input type="text" value="20.0"/>	<input type="text"/>
Microfilm	<input type="text" value="5.0"/>	<input type="text" value="7.0"/>	<input type="text" value="10.0"/>	<input type="text"/>
Diapositivas	<input type="text" value="1.0"/>	<input type="text" value="2.0"/>	<input type="text" value="3.0"/>	<input type="text"/>

Fig. 3. Parameters used to estimate digitization costs

Estimación de tiempos - CORRECCIÓN

Id: Id.Creador reg.: Nombre:

Núm.reg.: Título: Autor/es:

Número de pág. digitales: Dificultad lectura: Corrector ortográfico
 Número de palabras: Dificultad tablas: Aparato crítico
 Número de pág. papel: Dif. marcas y notas: Hiperenlaces-imágenes
 % Errores OmniPage: Formato publicación: Idiomas extranjeros
 Otras fuentes Teclado

Observaciones e incidencias:

Ajuste tiempo: 0:00
 T. estimado: 60:30
 T. almacenado: 60:30

Fig. 4. Estimate of correction costs

ESTIMACIÓN TIEMPO - CATALOGACIÓN		
Fecha inserción registro	30/05/2002 11:22:07	Ver estadillo
Registro creado por	espe	
Tiempo estimado (minutos)	30 0:30	

ESTIMACIÓN TIEMPO - DIGITALIZACIÓN		
Fecha inserción registro	08/07/2002 19:24:02	Ver estadillo
Registro creado por	icarlos	
Tiempo estimado (minutos)	129 2:09	

ESTIMACIÓN TIEMPO - CORRECCIÓN		
Fecha inserción registro	31/05/2002 10:14:55	Ver estadillo
Registro creado por	quica	
Tiempo estimado (minutos)	3630 60:30	

ESTIMACIÓN TIEMPO - VALOR AÑADIDO		
Fecha inserción registro		Ver estadillo
Registro creado por		
Tiempo estimado (minutos)		

ESTIMACIÓN TIEMPO - ED. FACSIMILAR		
Fecha inserción registro		Ver estadillo
Registro creado por		
Tiempo estimado (minutos)		

Fig. 5. Final report of digitization costs for a book (shows cataloging, scanning and correction of the text)

markup in the case of text production, and cataloging, scanning, and graphic processing in the case of facsimiles.

A few screenshots captured from this system are shown below. Figure 2 shows a scanning-only estimate for a 71 pages book. Figure 3 shows average historic values for different types of complexities and types of scanning device. Figure 4 shows a correction-only estimate, and Figure 5 shows the final summary of costs for the production of a digitized text book.

3 Conclusions

We have developed a cost estimation model for digitization projects based on known software engineering cost models. This method allowed us to predict the time required to complete digitization tasks with good accuracy. Digitization projects, compared to software development projects, have the advantage that the size of the work to be done can be known beforehand (namely the number of pages or words to digitize). In software design we can only guess the total number of lines of code a project will require, and the accuracy of the calculated time estimates will depend largely on this preliminary “expert judgment” estimate.

We verified that the model we propose works well in practice, and can be easily applied to different digital production processes, or other project or engineering tasks, provided that the cost equation is fine-tuned for each type of task using historical data. This requires two things to be done advance:

- Sufficient historical data must be collected to fine-tune the parameters of the cost equation.
- The main objective factors that affect the time required to do the task must be determined, and adequate effort adjusting modifiers be calculated and assigned to each of them.

With this information, a cost-equation for the specific production process can be easily obtained. Good expert knowledge of the process facilitates the fine-tuning task and allows for better estimation equations. Nevertheless, the cost-equations can be dynamically improved by re-adjusting the parameters with the new data fed-back from recently finished projects. In this way the estimation model can be continuously and incrementally improved.

3.1 Some Remarks on the Nature of Time and Cost Estimates

Often we use the words prediction and forecast when referring to estimates. The nature and purpose of predictions and forecasts is different from estimates. In the case of predictions and forecasts (think of stock-exchange predictions or weather forecasts), we obtain some prediction values, and then wait for real events to happen and confirm the predictions, or not. In the case of an estimate, we should not wait for an event to happen, but should work towards it instead. This active, not passive, nature is essential for profiting from estimates. An estimate is a target, a goal we have to fulfill, a reference or time frame to help us control our project. A good estimate is the time or cost objective withing which a task **can be done under moderate pressure** with a **reasonably good quality**. A task can always be done in a longer time, or in a shorter time under exceptional pressure, up to a point when either it cannot be done (at least with the required quality), or it produces undesired uneasiness in the work team. So it is wise to think of estimates as **reasonable goals**, that will require some effort and control, and not as mere predictions. A good deal of risk management is also advisable to help accomplishing the estimated targets, without surprises.

References

1. Albrecht, A.J., Gaffney, J.E.: Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation. IEEE Transactions on Software Engineering SE-9(6), 639–648 (1983)
2. Banerjee, G.: Use Case Points, An Estimation Approach (August 2001), http://www2.fiit.stuba.sk/bielik/courses/msi-slov/reporty/use_case_points.pdf
3. Boehm, B.W.: Software Engineering Economics. Prentice Hall, Englewood Cliffs (1981)

4. Boehm, B., Clark, B., Horowitz, E., Westland, C., Madachy, R., Selby, R.: Cost Models for Future Software Life-Cycle Processes: COCOMO 2.0. In: Arthur, J., Henry, S. (eds.) *Annals of Software Engineering Special Volume on Software Process and Product Measurement*, vol. 1, pp. 45–60. J.C. Baltzer AG, Science Publishers, Amsterdam (1995)
5. Clark, B., Devnani-Chulani, S., Boehm, B.: Calibrating the COCOMO II Post-Architecture Model. In: *20th International Conference on Software Engineering*, Center for Software Engineering, Computer Science Department, University of Southern California, Los Angeles, CA 90098-0781 USA, +1 213 740 6470 (April 1998), <http://sunset.usc.edu/csse/TECHRPTS/1998/usccse98-502/CalPostArch.Pdf>
6. CSE: COCOMO II Model Definition Manual, Center for Software Engineering, Computer Science Department, University of Southern California, Los Angeles, Ca. 90089 (1997), http://sunset.usc.edu/csse/research/COCOMOII/cocomo2000.0/CII_modelman2000.0.pdf
7. DeMarco, T., Lister, T.: *Peopleware, Productive Projects and Teams*. Dorset House Publishing, New York (1987)
8. Fairley, R.E.: *Software Engineering Concepts*. McGraw-Hill, New York (1985)
9. Galorath, D.: *Software Project Failure Costs Billions. Better Estimation and Planning Can Help*, June 7 (2008), <http://www.galorath.com/wp/software-project-failure-costs-billions-better-estimation-planning-can-help.php>
10. LCI: *Use Cases and Function Points*, Longstreet Consulting Inc. (2004), <http://www.ifpug.com/Articles/usecases.htm>
11. Magazinovic, A.: *Exploring Cost Estimation Inaccuracy - Why do practitioners still fail to predict the actuals?* Tech. rep., Chalmers University of Technology, Department of Computer Science and Engineering, Chalmers University of Technology, SE-41296 Göteborg, Sweden (2008), <http://publications.lib.chalmers.se/cpl/record/index.xhtml?pubid=73759>
12. Minkiewicz, A.F.: *Measuring Object Oriented Software with Predictive Object Points*, PRICE Systems, L.L.C (1997), http://www.pricystems.com/white_papers/Measuring%20object%20oriented%20Software%20with%20Predictive%20object%20Points%20July%20%27%20-%20Minikiewicz.pdf
13. Sackman, H., et al.: *Exploratory Experimental Studies comparing Online and Offline Programming Performance*. *Communications of the ACM* 11(1) (January 1968)

In Pursuit of an Expressive Vocabulary for Preserved New Media Art

Andrew McHugh and Leo Konstantelos

Humanities Advanced Technology and Information Institute
University of Glasgow, UK

a.mchugh@hatii.arts.gla.ac.uk,
l.konstantelos@hatii.arts.gla.ac.uk

Abstract. The status of the new media, interactive and performance art context appears to complicate our ability to follow conventional preservation approaches. Documentation of digital art materials has been determined to be an appropriate means of resolving associated difficulties, but this demands high levels of expressiveness to support the encapsulation of the myriad elements and qualities of content and context that may influence value and reproducibility. We discuss a proposed *Vocabulary for Preserved New Media Works*, a means of encapsulating the various information and material dimensions implicit within a work and required to ensure its ongoing availability.

1 Introduction

Numerous layers (both physical and conceptual) support encapsulation of and access to digital information, in contrast with analogue information, which is largely atomic. Within a new media creative context, Rinehart [7] expresses this layeral complexity in terms of the separability of the physical and the logical, which in turn creates opportunities for variation of behaviour and performance. It also limits self-evidence of such materials, and introduces difficulties from a preservation perspective.

For performative works and for complex interactive installations the most critical dimension of preservation role is not maintaining a work *per se*, but instead preserving sufficient information to facilitate its recreation at a later date, in a manner consistent with the original creative intention. This might appear detached from the preservation objectives of libraries or archives that focus largely on the object or record, and on maintaining its availability and authenticity throughout temporal and contextual change. But the difference is one of skewed emphases rather than substantively different priorities. More so than preserving a tangible thing, the real purpose is the preservation of the end user's experience, irrespective of material specificities.

Interestingly, despite the shared purpose that art conservators and curators share with library and archival communities, there is little evidence of cross-pollination of theory or practice. For instance, despite its exploration of many

aspects of information and bit-level preservation, the *Conservation Guide* published by the *Documentation and Conservation of the Media Arts Heritage Group* (DOCAM) contains no references to mainstream preservation literature [3]. Work in characterisation and preservation planning illustrated in Planets¹ with the eX-tensible Characterisation Language and the Plato [9] planning tool, and work in empirical evaluation supported by tools like the Planets Testbed [1] must find their applicability in this and other contexts. This demands a common, mutually applicable approach to the preservation challenge.

2 Previous Work

With the Media Art Notation System (MANS), Rinehart [7] acknowledges the performative characteristics of new media art materials, and seeks to conceive implementation agnostic means of describing materials' value. A noted shortcoming for preservation applications is MANS' association of *Descriptor* elements with each material *Resource*, intended to enable the explication of appropriate preservation strategy. However, this appears to prioritise physical aspects of preservation with less focus on the origins of particular information properties of value. The relationship between MANS' logical *Parts* and material *Resources* is not explicit, introducing difficulties in forming links between proposed preservation solutions (or, much more usefully, potential preservation risks).

The InSPECT project presents a workflow [5] aimed at the identification of significant properties, adopting a terminological foundation traceable to but distinct from MANS. Its FBS model (derived from Gero's *Function-Behaviour-Structure* Framework [4]) defines *Function* as broad purpose, *Behaviour* as a stakeholder's perceived outcome or consequence, and *Structure* as those elements of a given digital object that support a behaviour's realisation (significant properties). Stakeholder and object analyses demand and engagement with diverse stakeholders and identification of functional facets of value.

Within the National Archives of Australia's Performance model, also visible within OAIS' representation information concept [2], we synonymise software performance with data's associated process. Its application to a data source yields a data performance. A JISC Framework for Software Preservation, which followed on from an earlier study into significant properties of software [6] presents a four layer model for software as *Product*, *Version*, *Variant* and *Instance* that is roughly analogous to the FRBR model of Work, Expression, Manifestation and Items [8]. Applied to the new media context, process can be interpreted as having technological, procedural or semantic facets. As well as software, documentation remains an integral dimension irrespective of whether one is describing the required steps for collecting and arranging wooden sticks to reproduce a physical art installation² or encoding appropriate representation information to reveal the meaning of coded column headings in an Excel dataset.

¹ See <http://www.planets-project.eu/>

² See Meg Webster's "Stick Spiral" (1986).

3 Expressive Documentation for New Media Art

Our vocabulary is positioned firmly within the domain of new media art preservation. Instead of focusing on the description of materials in and of themselves we look to conceive a description of the *Preserved New Media Work*. This implies that some elements of preservation infrastructure become implicit within the work itself. While perhaps not part of the piece envisaged by the creator these become nevertheless integral to its ongoing survival, like a pacemaker inserted into a human heart. Naturally, as the artist's view takes on such critical importance within this domain, only those additions that have been satisfactorily sanctioned can occupy such a role.

The *Vocabulary for Preserved New Media Works* (VPNMW) collates a complex set of information that may relate to multiple individual instantiations of a work across space and time. Likewise it is sufficiently loosely defined to support additional variability within the process of preservation. We assume a number of relationships between its principle dimensions; *Work*, *Version*, *Functional* and *Material Component*, *Dependency*, *Context*, *Property* and *Stakeholder*.

- Our parent element is the **Preserved New Media Work**, encapsulating every intellectual and material facet of the preserved work. This includes both elements of the artistic work and constituents of the preservation process. **Stakeholders** are associated with the work and have a range of priority levels for legislating on the work's value components, and determining acceptable limits for the preservation and management process.
- Works have multiple **Functional Components**, consistent within a single work, and contributing largely to its definition. This does not imply that they are completely static, as through their relationship with variable **Properties** a range of acceptability is established.
- **Functional Components** can exist hierarchically, and therefore single functional behaviours' may be grouped into wider functions. Multiple **Versions** may exist within a single preserved new media work, a consequence of variability within the creative space, and also of preserved outputs, which may differ from the original. Different versions share function, but may exhibit material differences. Within the context of each version there must be an explicit mapping between **Material** and **Functional** elements. There must be assurances of sufficient materiality to satisfy functional and property requirements. Specific versions may benefit from input of alternative or additional **Stakeholders** to the work itself; therefore, versions can be related directly to individual stakeholders.
- **Material Components** are the tangible building blocks of the preserved work, considered distinct from function, but directly contributing to its realisation. The relationship between material and function can be 1:1, but likewise in any single version there may be several material assets associable with a single function, and by extension significant properties. Any additional

documentation assumes the character of material component, and becomes part of the PNMW.

- **Dependencies** are any process elements that must be associated with material elements to realise functional or property requirements of a work. These may be procedural, or infrastructural, or based on particular contextual qualities. At times it may be necessary to absorb **Contextual** elements into a work as an integrated dependency in order to resolve contextual omissions that occur over time (e.g. to provide an audience with the understanding that a worldwide recession took place at the end of the first decade of the 21st century). Dependencies are representation information; they may be structural or semantic, but are integral to establishing functional sense from material components.
- **Context** describes factors that exist outwith the control of the preservation environment, but that contribute to either its function (and associated properties) or are required as dependencies to realise material component's performance. Context is a critical dimension for documentation, since it cannot be manipulated directly by the preservation professionals. There is scope to absorb evidence of contextual elements into the *PNMW* as documentation, and these are encapsulated as material components.
- **Stakeholders** are the individuals that perform the preservation activity. Among their primary goals are to determine functional components (and by association properties) and their acceptable variability; evaluate material version-specific components to ensure their capacity to satisfy functional component requirements; monitor dependencies and contextual circumstances to ensure their ongoing adequacy; evaluate preservation risks and conceive, exercise and validate appropriate preservation responses.
- **Properties** are those measurable facets of function that collectively express the value of a *PNMW*, and that must persist for preservation to succeed. Properties are related to a stakeholder who explicates their identity and value, as well as (crucially) an acceptable tolerance for variability. Properties are frequently associated with function, but can also relate to dependencies and context as a means of expressing acceptable variability that can be tolerated before a preservation interaction is required.

3.1 Preserved New Media Work

At the top-most level of our information infrastructure we have the concept of a *Preserved New Media Work*. This has a number of sub-dimensions, which must be related and rationalised in order to determine preservation challenges and equip ourselves to satisfy them appropriately. It is at this top level that we associate descriptive metadata information, and other registration details that describe the work as a whole. There is value in presenting this information at the level of work, although further granularisation at the level of individual components and contextual elements enables more sophisticated and finely tuned recording, and associated preservation planning.

3.2 Functional and Material Components

A critical foundation for supporting works' recreation are means to describe both the intellectual object of preservation, and those physical material manifestation of that information. Content within a new media art piece may be as potentially diverse as one could possibly envisage, including real world objects, digital media, and combinations of both. More critical than considering objects in tangible terms is their expression as measurable (and functional) properties, ideally in a manner that is agnostic to any transitory, non-specific implementation. MANS elects to approach preservation as an activity that practically focuses on tangible system components (Resources), with an expectation that their preservation will safeguard the more intellectually (or functionally) specific Parts. This seems short-sighted we need not retain physical equivalence to ensure the sustainability of logical meaning. For example, it may be possible to replace multiple discrete media assets (e.g. still images, sound materials, interview transcripts) with a single subtitled video and retain every aspect of original information value. The message is the critical point at which persistence must be sought the physical building blocks are merely means to that end. This is why documentation can occupy a partial-surrogacy role, and be capable of expressing aspects of original meaning.

Even where artists stipulate conditions that appear to concern only matters of physicality, we must interpret that in intellectual terms. If a particular model of display device must be used for example we must consider that in its functional terms (i.e., its creative significance), rather than interpreting it as a material requirement. We should not assume a 1:1 correspondence between material and intellectual components.

The functional component is best expressed in terms of properties; this affords a level of measurability that is required to validate preservation efforts, and to make explicit acceptable boundaries for variability which are an intrinsic part of especially these kinds of materials.

New media works are dynamic and therefore may have multiple manifestations available simultaneously or along a time line. The version element provides a means to accommodate this dynamic quality, with the potential for multiple instances of a work which while tangibly variable nevertheless represent the same conceptual piece. Although material aspects of the work may vary across versions the functional components (expressed primarily in terms of associated, and a bounded range of property values) will remain consistent.

A complication facing the preservation community is that factors threatening our information often do not do so directly. Although the preservation goal is targeted on the sustainability of more intellectual or functional facets, it is often tangible and physical characteristics that are threatened by specific preservation risks (for example, the risk of file format obsolescence). This is not uniformly true we also face challenges such as insufficiency of semantic representation information for example, but the disconnect demands an understanding of the interrelationships between each dimension. We distinguish a work's functional and material character to support better preservation decision making. Material

components are intended to encapsulate a physical, and, one would anticipate, transitory dimension of a work. Their availability is threatened by preservation risk, which demands our awareness of the relationship between risk and materiality. Having established such links, of greatest importance is their relationship with intellectual properties, and by extension function.

3.3 Component Dependency

Both material and functional components exhibit dependencies, and again we must make this relationship explicit within our vocabulary. Dependencies describe those facets of process that must exist to support the realisation, from a content source, of an information performance. These may assume myriad forms, including technical or other infrastructural (most obviously software), procedural or contextual dependencies. Once more, these dependencies are expressed at the level of a preserved work, meaning that there are a number of examples included primarily due to the role they perform within the preservation process.

3.4 Work Context

The primary purpose of recording contextual dimensions is to make explicit those external or situational influences that must persist or be recreatable to realise or perform a work and preserve original artistic intention. Context is distinct from implicit components, dependencies and stakeholder relationships, in that they may surround, influence and reflect either the global work (or in even wider terms whole collections) or just individual information facets. Many contextual facets are represented as points on a continuum – variability and evolution of a work implies movement along this continuum, and reflects the different contextual properties that may still surround and legitimise a work.

Context is distinct from content in terms of the extent to which it can be realistically preserved. We cannot hope to maintain every aspect of context. In some respects objects and their associated representation mechanisms may exhibit change over time (for example, in the case of bit-rot), but the greatest challenge for preservation professionals is keeping up with change that is wholly contextual, whether realised in financial, technological or cultural terms, almost always a reactive process.

Preservation requires the establishment (probably with the input of artists) of acceptable spectrums for contextual deviation. For example, what spatial restrictions are tolerable on a particular installed piece? What opportunities are there to transfer content to new media devices? What wider contextual factors (for example a financial recession) must be documented and integrated within a work to maintain its essence when those factors have since changed and been forgotten? In these respects the line between context and content (particularly objects' associated dependencies or process elements) may appear blurred; the preservation process demands the explication of that which is content, and that which is a relevant, but not integral contextual factor. Likewise, for each contributing factor, tolerable parameters and descriptions of associated documentation requirements should be made explicit.

3.5 Stakeholder

The diversity of roles and priorities that contribute to the creation, documentation, preservation and consumption of art hints at the complexity of the characterisation process. Artists are most naturally assumed to be the most appropriate arbiter of a work's significance. Likewise, they are often relied upon to sanction preservation interventions that may otherwise prejudice its value. Example accounts exist of useful artist intervention [10], but this probably cannot be expected to be typical. Nevertheless, engagement with creators is a critical part of understanding the work, and the breadth of opportunities for its preservation.

The other broad dimension of stakeholder intervention is identification of preservation risk and challenge. For bespoke highly complex technical materials this may presuppose the input of wider constituencies than simply curators. Technological contributors for example are very well placed to comment on information dependencies implicit within any code they have implemented for a specific work. Curators must assume primary responsibility for preservation risk awareness, although as described above this assumes a close understanding of the relationships between a work's tangible assets and softer facets of message and value, expressed as properties.

3.6 Information Property

Preservation planning must be moored to both the tangible realities of a piece and their cumulatively realised expressive force. This softer, but most critical dimension is best expressed in terms of properties. Information properties are the focus of the preservation effort, and are potentially limitlessly diverse. Each specific property has a number of individual facets. They are relatable to both functional and material components, and to stakeholders, who are at least partially responsible for their definition, and for establishing bounds of acceptability for variation of those properties over time.

A well defined information property is one that is discrete, measurable and explicit. There are few if any information domains where such attributes are universally feasible. There are always likely to be peripheral, but nevertheless potentially integral properties that are inarticulately defined, or insufficiently tangible to express in empirically evaluable terms. A pragmatic approach may be to ignore these in favour of those properties that can be definitively validated (ideally using automated tools) but this remains unsatisfactory, particularly for qualities (frequently associated with new media art) that are ephemeral or philosophical. The primary role of new media art preservation and documentation is to distil even loosely expressed properties into tangible factors that can be exposed to validation. The characterisation process must seek to granularise works into discrete component parts, each composed of some kind of content, associated dependencies, implicit variability, and stakeholder relationships. These are then further subdivided into associated properties, and aligned with a characterisation of causally or effectually linked context.

4 Conclusion and Further Work

This short paper introduces a possible vocabulary for supporting new media art preservation, building on foundations established in preservation research in both creative and more mainstream information domains. Future work will seek to implement the vocabulary as an ontology and validate its effectiveness in real-world new media conservation and curation environments.

Acknowledgments

This work was first conceptualized in and supported by the *Planets* (IST-2006-033789) Project, funded by the European Commission's IS&T 6th Framework Programme.

References

1. Aitken, B., Helwig, P., Jackson, A., Lindley, A., Nicchiarelli, E., Ross, S.: The planets testbed: Science for digital preservation (2008)
2. Consultative Committee for Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS) (January 2002), ISO 14721:2003
3. Documentation and Conservation of the Media Arts Heritage. Conservation Guide (2009)
4. Gero, J.: Design prototypes: A knowledge representation schema for design. *AI Magazine* 11(4), 26–36 (1990)
5. Knight, G.: Framework for the definition of significant properties (2008)
6. Matthews, B., Bicarregui, J., Shaon, A., Jones, C.: Framework for software preservation. In: Proc. 5th International Digital Curation Conference (IDCC 2009), London, UK (2009)
7. Rinehart, R.: A system of formal notation for scoring works of digital and variable media art. In: *Archiving the Avant Garde* (2005)
8. Saur, K.: Functional requirements for bibliographic records: final report / IFLA Study Group on the Functional Requirements for Bibliographic Records, Munich (1998)
9. Strodl, S., Becker, C., Neumayer, R., Rauber, A.: How to choose a digital preservation strategy: evaluating a preservation planning procedure, pp. 29–38 (2007)
10. Hummelen, Y., Sille, D. (eds.): *Modern Art: Who Cares? An interdisciplinary research project and an international symposium on the conservation of modern and contemporary art*. Stichting Behoud Modern Kunst/Instituut Collectie Nederland, Amsterdam (1999)

Privacy-Aware Folksonomies

Clemens Heidinger, Erik Buchmann, Matthias Huber,
Klemens Böhm, and Jörn Müller-Quade

Karlsruhe Institute of Technology, Germany
{heidinger,erik.buchmann,matthias.huber,
klemens.boehm,mueller-quade}@kit.edu

Abstract. Many popular web sites use folksonomies to let people label objects like images (Flickr), music (Last.fm), or URLs (Delicious) with schema-free tags. Folksonomies may reveal personal information. For example, tags can contain sensitive information, the set of tagged objects might disclose interests, etc. While many users call for sophisticated privacy mechanisms, current folksonomy systems provide coarse mechanisms at most, and the system provider has access to all information. This paper proposes a privacy-aware folksonomy system. Our approach consists of a partitioning scheme that distributes the folksonomy data among four providers and makes use of encryption. A key sharing mechanism allows a user to control which party is able to access which data item she has generated. We prove that our approach generates folksonomy databases that are indistinguishable from databases consisting of random tuples.

1 Introduction

Folksonomies [18] are a popular Web 2.0 approach to let a community of users annotate large sets of objects with schema-free labels (tags). For example, Last.fm lets its users tag music, Flickr uses a folksonomy to annotate photos, and Delicious builds a directory of web pages based on a folksonomy. Users generate tags for various reasons [2], e.g., to share information with friends, to find people with similar interests, or to organize objects for personal use.

Folksonomies may contain personal information. Since the users are free to annotate any object with arbitrary tags, folksonomy data can reflect interests, habits, behavior, social aspects, and other sensitive characteristics. For example, Delicious discloses the web pages a user has tagged, the time and day she generated the tags, her attitude towards the web pages, and the network of users with similar interests. The privacy controls being part of most folksonomy systems are coarse at best. For example, Flickr considers tags to be private if the tagged photos are private. Furthermore, with all folksonomies we are aware of, the provider has access to any user-generated data. However, most users wish to control in a fine-grained way who can access which personal information [5].

In this work, we propose an approach towards privacy-aware folksonomies. This is challenging: A privacy-aware folksonomy has to support individual privacy preferences. Each user must be able to allow or forbid other users to access

the data she has generated. For example, a user might wish to share her tags with anybody, but she might want to share the tagged objects with friends only. Furthermore, we have to consider that folksonomy systems process many parallel queries from different users on large data sets. Thus, expensive approaches like secure multiparty computation [7,10] are not practical.

While guaranteeing privacy under relatively weak assumptions, our approach is simpler, as follows: It partitions folksonomy data into four databases. Each database is encrypted and stored at a different provider. A key-sharing mechanism allows the users to control who is able to access which data they have generated. Our approach produces folksonomy databases that are indistinguishable from a perfectly anonymized database, i.e., a database that is indistinguishable from a database with random tuples. We provide a proof sketch that our approach ensures data privacy according to k -Ind-ICP with ϵ -Advantage, a notation that maps probabilistic encryption to k -Anonymity [20].

Summing up, we make the following contributions:

- We develop privacy requirements for folksonomies by analyzing operations on popular folksonomies and characteristics of the data stored.
- We present our approach for privacy-aware folksonomy systems.
- We evaluate our approach against k -Ind-ICP with ϵ -Advantage.

Paper outline: Section 2 reviews related work. Section 3 introduces folksonomies and privacy threats. Our privacy-aware folksonomy is described in Section 4. Section 5 evaluates this approach against a security model, and Section 6 concludes.

2 Related Work

We consider three classes of related work: (1) Folksonomies and respective privacy issues, (2) privacy approaches, and (3) formal definitions of privacy.

Folksonomies [18] are used in many web applications. [5] investigates the privacy preferences of the users of a geo-tagging folksonomy. The users call for sophisticated privacy mechanisms to keep 12% of all tags and 14% of all geo-coordinates private, and they distinguish between different social groups when sharing private information. It is important to know what motivates people to tag, and which kind of tags are commonly used. [2] shows that there are (i) social and (ii) functional motivations. People generate tags for themselves, for various social groups, and for the public good. Regarding function, [2] distinguishes between organization and communication. [13] has categorized tags. Affective tags (e.g., “cool”, “fun”, or “dull”) that fall under the category “non-subjective tags” are most privacy-relevant.

Approaches for secure and private databases are related as well. [12] proposes an encryption scheme for databases stored at untrusted parties. The approach has a security risk: Query processing requires the database server to know encryption keys. [11] proposes an encryption scheme to execute SQL on encrypted data. [6,8,14] refine this concept. [1] vertically partitions a database among two providers according to privacy constraints. However, the approach leaves associations between deterministically encrypted attributes intact. As it is known that

the frequency of folksonomy data usually follows a power-law distribution [18], this approach would allow statistical attacks if applied to a folksonomy. [15] implement access control for annotation services. However, all data is stored unencrypted at one provider.

We verify our approach against a security notation that is related to k -Anonymity [20]. A database is k -anonymous if for each entry that refers to an individual, there are at least $k - 1$ other entries that can be mapped to the same individual. l -Diversity [17] and t -Closeness [16] extend k -Anonymity. Nevertheless, as the majority of folksonomy users generate only few tag applications [18], k -Anonymity and similar approaches require to delete or generalize a large number of tag applications, for any meaningful value of k . Our security notion k -Ind-ICP avoids this problem by mapping probabilistic encryption on k -Anonymity.

3 Folksonomies and Privacy

In this section, we provide a formal definition of folksonomies, we describe popular operations on folksonomy data, and we introduce our requirements.

3.1 A Formal Folksonomy Model

We model a folksonomy as a set of tuples F . Each tuple (o, u, t, d) consists of four attributes object o , user u , tag t , and creation date d . We refer to such a tuple as *tag application*. Note that other meta-data besides the creation date can be stored as well. Each attribute can carry sensitive information. For example, an object can be a photo that shows persons, or a tag can contain personal details like a name. Furthermore, operations on folksonomy data can reveal personal information. For instance, a tag cloud [19] compiled from all data provided by one person can reveal her interests. Our objective is to let each user control who can access the folksonomy data she has generated, and to find out which folksonomy operations are affected if parts of a folksonomy database cannot be accessed. We have analyzed the APIs and user interfaces of Flickr and Delicious, and we have identified nine popular operations:

F1 All tags a user u has applied:

$$f_1(u) = \{ \hat{t} \mid \exists \hat{o} \exists \hat{d} : (\hat{o}, u, \hat{t}, \hat{d}) \in F \}$$

F2 All objects a user u has tagged:

$$f_2(u) = \{ \hat{o} \mid \exists \hat{t} \exists \hat{d} : (\hat{o}, u, \hat{t}, \hat{d}) \in F \}$$

F3 All tags a user u has assigned to an object o :

$$f_3(u, o) = \{ \hat{t} \mid \exists \hat{d} : (o, u, \hat{t}, \hat{d}) \in F \}$$

F4 All objects a user u has assigned with a tag t :

$$f_4(u, t) = \{ \hat{o} \mid \exists \hat{d} : (\hat{o}, u, t, \hat{d}) \in F \}$$

F5 Tag cloud for object o :

$$f_5(o) = \{ \hat{t} \mid \exists \hat{u} \exists \hat{d} : (o, \hat{u}, \hat{t}, \hat{d}) \in F \}$$

F6 Objects recently tagged (d is a date in the past):

$$f_6(d) = \{ \hat{o} \mid \exists \hat{u} \exists \hat{t} \exists \hat{d} : (\hat{o}, \hat{u}, \hat{t}, \hat{d}) \in F \wedge \hat{d} \geq d \}$$

- F7** Tags recently used (d is a date in the past):
 $f_7(d) = \{ \hat{t} \mid \exists \hat{o} \exists \hat{u} \exists \hat{d} : (\hat{o}, \hat{u}, \hat{t}, \hat{d}) \in F \wedge \hat{d} \geq d \}$
- F8** Objects assigned with a specific tag t :
 $f_8(t) = \{ \hat{o} \mid \exists \hat{u} \exists \hat{d} : (\hat{o}, \hat{u}, t, \hat{d}) \in F \}$
- F9** Tags assigned to a specific object o :
 $f_9(o) = \{ \hat{t} \mid \exists \hat{u} \exists \hat{d} : (o, \hat{u}, \hat{t}, \hat{d}) \in F \}$

Since the computation of a tag cloud requires to know the number of identical tags, **F5** returns a multi-set. Clearly, further operations are conceivable, because folksonomies can be used for many different applications. However, this list of operations serves as a reference for popular methods on folksonomy data. Table 1 shows the operations implemented (✓) by Flickr and Delicious.

Table 1. Operations implemented by Flickr and Delicious

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
Flickr	✓	●	✓	✓	●	✓	✓	✓	✓
Delicious	✓	✓	✓	✓	✓	✓	●	✓	✓

3.2 Requirements for a Privacy-Aware Folksonomy System

All popular folksonomy systems we are aware of provide only coarse privacy controls. For example, Delicious allows to make objects and tags invisible for anybody except the creator. The privacy controls of Flickr are similar, but also distinguish between friends and relatives. However, as we know from previous studies, e.g., [5], folksonomy users wish to define who is allowed to access which personal information in a fine-grained way. Neither the folksonomy provider nor the users should be able to process data without authorization of its creator. Finally, as folksonomy systems store huge data sets, a privacy-aware folksonomy must be scalable. We have compiled the following requirements:

- R1: Limited disclosure.** The folksonomy provider should learn as little information as possible when a user generates a tag application.
- R2: Operation support.** Folksonomy users must be able to control who can access data they created.
- R3: Performance.** The privacy mechanisms must be scalable in order to not limit the performance of the folksonomy system.

These requirements conflict with each other. However, folksonomies usually do not contain highly sensitive attributes such as the social security number, which would require strong cryptography by all means. Thus, we strive for a reasonable tradeoff between **R1**, **R2**, and **R3**.

4 A Privacy-Aware Folksonomy System

In this section we describe our approach towards privacy-aware folksonomies. We propose a *separation of duties* approach for a privacy-aware folksonomy system. It distributes the folksonomy database among four independent providers and uses

encryption. The result is a distributed database that is indistinguishable from random data, as we will show in Section 5.2. In the following, we will describe the details of our approach, and we will show that it meets our requirements.

4.1 Separation of Duties

Folksonomies store (user, tag, object)-tuples and a creation date. To prevent that a provider can observe if a particular user generates a tag application, we store those attributes in four databases $\text{index}_{\text{user}}$, $\text{index}_{\text{tag}}$, $\text{index}_{\text{object}}$ and *associations*. The index databases store tuples consisting of a pointer and a value. The value represents a user, tag or object, and each index database is responsible for one of these attributes. Each tuple in the associations database contains three pointers from the index databases, the tag application and the creation date. Each database is stored at a different provider.

Next, we propose the following encryption scheme: The values in the index databases are encrypted with a deterministic encryption enc_d . Deterministic encryption always produces the same ciphertext for a given plaintext and a given key. This enables the provider to efficiently search for encrypted values in sublinear time [3] without knowing the plaintext. The pointers in the index databases and the tag applications in the associations database are encrypted with a probabilistic encryption enc_p . A probabilistic encryption results in a different ciphertext with a high probability when encrypting the same plaintext multiple times with the same key.

Probabilistic encryption is needed for the encrypted databases to be indistinguishable from databases containing the same number of random values, even

Table 2. Plain-text folksonomy

User	Tag	Object	Time
Alice	toread	Blog	2010-01-15 15:23
Alice	towatch	Video 1	2009-08-14 22:11
Bob	towatch	Video 1	2010-01-27 09:14
Bob	towatch	Video 2	2010-01-27 09:17

Table 3. Distributed and encrypted folksonomy

(a)

Provider U: $\text{index}_{\text{user}}$

value	pointer
$\text{enc}_d(\text{Alice})$	$\text{enc}_p(r_1, r_2)$
$\text{enc}_d(\text{Bob})$	$\text{enc}_p(r_3, r_4)$

(b) Provider T: $\text{index}_{\text{tag}}$

value	pointer
$\text{enc}_d(\text{toread})$	$\text{enc}_p(r_5)$
$\text{enc}_d(\text{towatch})$	$\text{enc}_p(r_6, r_7, r_8)$

(c) Provider O: $\text{index}_{\text{object}}$

value	pointer
$\text{enc}_d(\text{Blog})$	$\text{enc}_p(r_9)$
$\text{enc}_d(\text{Video 1})$	$\text{enc}_p(r_{10}, r_{11})$
$\text{enc}_d(\text{Video 2})$	$\text{enc}_p(r_{12})$

(d) Provider A: associations

P _u	P _t	P _o	User	Tag	Object	Time
r_1	r_5	r_9	$\text{enc}_p(\text{Alice})$	$\text{enc}_p(\text{toread})$	$\text{enc}_p(\text{Blog})$	$\text{enc}_p(2010-01-15 15:23)$
r_2	r_6	r_{10}	$\text{enc}_p(\text{Alice})$	$\text{enc}_p(\text{towatch})$	$\text{enc}_p(\text{Video 1})$	$\text{enc}_p(2009-08-14 22:11)$
r_3	r_7	r_{11}	$\text{enc}_p(\text{Bob})$	$\text{enc}_p(\text{towatch})$	$\text{enc}_p(\text{Video 1})$	$\text{enc}_p(2010-01-27 09:14)$
r_4	r_8	r_{12}	$\text{enc}_p(\text{Bob})$	$\text{enc}_p(\text{towatch})$	$\text{enc}_p(\text{Video 2})$	$\text{enc}_p(2010-01-27 09:17)$

if identical tags, users and objects appear many times in the folksonomy. In Section 5.2 we will prove this feature.

Example 1. Table 2 presents a folksonomy with four tag applications. Table 3 shows how our approach encrypts and distributes this folksonomy. \square

Listing 1 shows the pseudocode for a user generating a tag application. $enc_d(x)$, $dec_d(x)$, $enc_p(x)$ and $dec_p(x)$ encrypt and decrypt x with deterministic or probabilistic encryption respectively, and $random()$ generates a random pointer. $sendRecord()$ and $getRecord()$ send or receive information to/from a provider. Generating a tag application is processed as follows: For each attribute “user”, “tag” and “object” the user creates a random pointer (Line 2) and executes a lookup on the respective index database (Line 3). If the index database already contains a tuple for the encrypted attribute, the user decrypts the set of pointers, appends the new pointer, encrypts the pointer set and updates the tuple in the index database (Lines 7-9). If such a tuple does not exist, the user creates a new one that consists of the encrypted (attribute, pointer)-pair (Line 5). Finally, the user inserts the encrypted pointers and the encrypted tag application into the associations database (Line 12). Each provider only learns that a tag application was generated.

```

1  foreach a ∈ (user, tag, object) {
2      pointer pa = random();
3      record r = indexa.getRecord(encd(a));
4      if r = null {
5          indexa.sendRecord(encd(a), encp(pa));
6      } else {
7          pointerset l = decp(r.pointer);
8          l.append(pa);
9          indexa.sendRecord(encd(a), encp(l));
10     }
11 }
12 associations.sendRecord(puser, ptag, pobject, encp(user), encp(tag), encp(object),
    encp(date));
    
```

Listing 1. Performing a tag application

Now assume a user u wants to execute a folksonomy operation, e.g., **F2**: all objects u has tagged. To process **F2**, the user needs the keys for the deterministic encryption of u and the decryption of the objects she has tagged. First, she queries Provider U with $enc_d(u)$ to obtain a set of encrypted pointers to the tag applications containing u . The user then decrypts those pointers and sends them to Provider A. Provider A returns the encrypted tag applications, and the user can decrypt the objects.

4.2 Key Sharing

The encryption keys determine who is allowed to access and to link which data. Each component of a tag application, i.e., user, tag and object, can be encrypted with a different key according to the privacy preferences of the user. In this subsection, we propose a key-sharing scheme which allows to define who is allowed

Table 4. Key-sharing configuration of Alice

	User	Tag	Object
Self	✓	✓	✓
Friends	●	✓	✓
Provider	●	●	●

(a) Key sharing

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
Self	→	→	→	→	→	→	→	→	→
Friends	⊗	⊗	⊗	⊗	→	→	→	→	→
Provider	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗

(b) Executable operations

to access which data for the computation of folksonomy operations. We propose to use existing key-sharing services, e.g. [4], as technical infrastructure for key distribution.

The deterministic encryption enc_d prevents the provider of an index database from learning which users, tags or objects are referenced and queried. We propose that a folksonomy user deterministically encrypts all values in the index databases with the same key and shares it with all individuals that are allowed to access any data the user has generated. This reduces the overhead for key sharing. An individual who knows this key can only find out if the index databases contain a certain user, object or tag, but cannot decrypt the pointers to the associations database.

The probabilistic encryption enc_p secures tag applications in the associations database, and it conceals the pointers required to link the index databases to the associations database. Since computing folksonomy operations requires querying different index databases and different attributes from the associations database, the users can decide individually who may obtain which information from the folksonomy by (not) sharing the keys for the probabilistic encryption. In the following, we assume that each individual uses only one key for the deterministic encryption, and we describe two key sharing examples for the probabilistic encryption.

Example 2. Alice generates three different keys and encrypts each pointer in the index database and the respective attribute in the associations database with the same key, i.e., “pointer” in $index_{user}$ and “User” in associations are encrypted with the same key. Assume Alice does not want the providers to learn anything about her tag applications. She also does not want her friends to find out which tags she has applied and which objects she has tagged. Thus, Alice shares her keys as shown in Table 4. She knows all keys, her friends know the keys to decrypt “Tag” and “Object”, and the providers do not know any key.

Table 4 shows which party can (→) or cannot (⊗) process which operation on the data Alice has generated, given this key-sharing scheme. □

Example 3. Now consider Bob, a user who wants to keep the objects he has tagged secret. Furthermore, Bob does not want the providers to know his tag applications. Bob generates three different keys and shares them as shown in Table 5. Table 5 shows who can execute which operation on the data Bob has provided. □

Note that these two examples do not describe all configurations possible. In principle, each user, tag, or object can be encrypted with a different key. To mention a further example, assume the provider is given access to the keys

Table 5. Key-sharing configuration of Bob

	User	Tag	Object
Self	✓	✓	✓
Friends	✓	✓	●
Provider	●	●	●

(a) Key sharing

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
Self	→	→	→	→	→	→	→	→	→
Friends	→	⊗	⊗	⊗	⊗	⊗	→	⊗	⊗
Provider	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗

(b) Executable operations

required to decrypt object and time information in the associations database. Now the provider can identify objects recently tagged (**F6**), but cannot link them to a particular tag or user.

4.3 Performance Considerations

Our approach introduces computational overhead for encryption and network costs due to the distribution of the databases. In this subsection we will discuss the performance of our approach, and we will propose possible optimizations. Operations **F1-F4** and **F8-F9** require the following processing steps:

1. Encrypt one or more attributes with deterministic encryption *enc_d*.
2. Query index databases to obtain the probabilistically encrypted pointers.
3. Decrypt the sets of pointers returned.
4. Query the database “associations” with the pointers from the previous step to obtain (parts of) the probabilistically encrypted tag applications.
5. Decrypt the tag applications returned.

Both deterministic and probabilistic encryption and decryption can be processed efficiently on modern hardware. The search time on the index databases is sub-linear. In particular, due to the deterministic encryption, searching the database does not require to decrypt data, and the database can be optimized using indexes. The query on the associations database is a simple selection that can be done in logarithmic time. Thus, Steps 1-5 can be executed efficiently.

F5 is required to compute a tag cloud for a particular object. This operation returns a multi-set of tags, which can be very large in popular folksonomies. In traditional folksonomies, this operation can be implemented efficiently at a database server. With our approach, a large set of encrypted tags has to be transferred to and decrypted by the user who wants to build the tag cloud, if the provider does not know the keys required. However, we can optimize this operation by introducing a further database. It stores deterministically encrypted (object, tag)-pairs together with a probabilistically encrypted counter. The counter allows to determine the frequency of tags without transferring and decrypting a large number of them. It has to be updated if tag applications are inserted, updated or deleted. Table 6 provides an example for this optimization. This database does not comply with our security notion *k*-Ind-ICP as it allows to correlate the two deterministically encrypted attributes object and tag. Yet, attacks can only reveal if a certain tag was attached to a certain object, if the attacker knows the ciphertexts for tag and object.

Table 6. Performance optimization for computing tag clouds

Object	Tag	Counter
$enc_d(\text{Blog})$	$enc_d(\text{toread})$	$enc_p(1)$
$enc_d(\text{Video 1})$	$enc_d(\text{towatch})$	$enc_p(2)$
$enc_d(\text{Video 2})$	$enc_d(\text{towatch})$	$enc_p(1)$

Since the creation date in the “associations” database is encrypted, **F6** and **F7** require Provider A to return all encrypted objects and tags to the user. These operations can be optimized if Provider A is allowed to store the unencrypted creation date. As the provider can observe when the database is updated anyhow, this optimization does not reveal much additional information.

5 Security of Our Privacy-Aware Folksonomy

In this section, we briefly introduce our security notion k -Ind-ICP with ϵ -Advantage, and we prove that our approach is in line with this notion. An extensive discussion of our notion can be found in a complementary technical report [1].

5.1 Indistinguishability under Independent Column Permutation

A database cannot identify an individual, if it is indistinguishable from a perfectly anonymized database. Intuitively, a database is perfectly anonymized, if an adversary cannot distinguish between the anonymized database and a database that has been anonymized after all entries in each column of the original database have been permuted independently from each other. This holds because the permutation eliminates the relation between individual-related attributes. In order to provide different levels of privacy we adapt from k -Anonymity [20] and restrict the permutations to k rows. However, in this paper k is equal to size of the database n . Let f be our anonymization function. To show that f is a good anonymization, we specify an experiment where an adversary has to decide if an anonymized database has been permuted before the anonymization. Therefore, we define the following notions:

Definition 1. A database function is a function $g : DB \rightarrow DB$. We call \mathcal{D} the set of all database functions.

Definition 2. A quasi-identifier is a set of attributes that can be linked with external data to uniquely identify individuals (cf. [20]).

Definition 3. A column independent permutation $\mathbf{p} \in \mathcal{D}$ is a database function $\mathbf{p} : DB \rightarrow DB$ that permutes entries within each quasi-identifier column of a database but leaves other columns untouched. We call the set of all column independent permutations Π .

¹ http://sdqweb.ipd.kit.edu/huber/reports/sod/technical_report_sod.pdf

Examples for database functions are projection, selection, permutations Π , or their anonymization function f . Now we can define our security notion:

Definition 4. *Experiment $Ind-ICP_{\mathcal{A}}^{k,p,i}(d)$*

Let \mathcal{A} be a polynomially restricted adversary², $d \in DB$ be a database, and $i \in \{0, 1\}$. For each row r_j in a database d exists a set M_j of k rows in d (called the k -bucket of row r_j). Each $r_j \in M_j$ and each $\mathbf{p} \in \Pi$ affects only rows in M_j . All other rows remain unchanged. The experiment $Ind-ICP_{\mathcal{A}}^{k,p,i}(d)$ is:

$d_0 := f(d)$
 $d_1 := f(\mathbf{p}(d))$
 $b := \mathcal{A}(d_i)$
 return b

In experiment $Ind-ICP_{\mathcal{A}}^{k,p,0}(d)$, the adversary \mathcal{A} obtains an anonymization of the database d . In $Ind-ICP_{\mathcal{A}}^{k,p,1}(d)$, \mathcal{A} obtains an anonymization of a permutation of d . The advantage of \mathcal{A} is the advantage to guessing if the database was permuted before anonymization or not:

Definition 5. *Advantage of Adversary \mathcal{A}*

$$Adv_{\mathcal{A}}^{Ind-ICP^{k,p}}(d) := \left| Pr[Ind-ICP_{\mathcal{A}}^{k,p,0}(d) = 1] - Pr[Ind-ICP_{\mathcal{A}}^{k,p,1}(d) = 1] \right|$$

If the advantage of an adversary is smaller than ϵ , k -Indistinguishability under Independent Column Permutation holds:

Definition 6. *(k -Ind-ICP) with ϵ -Advantage*

For a database function f , k -Indistinguishability under Independent Column Permutation (k -Ind-ICP) with ϵ -Advantage holds iff for each polynomially restricted adversary \mathcal{A} , for each database $d \in DB$ the following holds:

$$Adv_{\mathcal{A}}^{Ind-ICP^{k,p}}(d) < \epsilon$$

For the sake of readability, we write “ k -Ind-ICP” instead of “ k -Ind-ICP with ϵ -advantage” whenever ϵ is reasonably small.

5.2 Proof Sketch

Let d be any database complying with the schema of Table 2. We first show that the function f_1 mapping d to an indexing database (e.g. the database in Table 3a) is k -Ind-ICP. The function f_1 suppresses all attributes except a given attribute a and encrypts the values of a with a deterministic encryption enc_d . Before encryption, f_1 either removes duplicate rows or uses different keys for encryption of these rows. In order to obtain the pointers for the association, f_1 adds a column containing a unique random number for each occurrence of

² The assumption of polynomially restricted adversaries is common in cryptography. An unbounded adversary with unlimited computing time, storage, etc. can break encryption schemes that rely on hard problems such as prime factorization.

the corresponding value of a in the original database. Further, f_1 encrypts all entries in the added column with a probabilistic encryption algorithm enc_p . The database $f_1(d)$ contains two columns. In the first column, no two values are identical, since f_1 removed duplicate rows or used a different encryption key. The values in the second row are random. So the table is indistinguishable from a table with mere random entries of the same size and hence f_1 is k -Ind-ICP for k being the complete number of rows of the database d , because permutations of entries in each column cannot be detected after application of f_1 .

Second we look at the association server. Let f_2 be a function that replaces every attribute value of a given database with a unique random number and additionally f_2 adds for each attribute a column containing original attribute values encrypted with a secure probabilistic encryption enc_p . The mapping f_2 generates an associations database exactly as in Table 3d. f_2 is k -Ind-ICP for k being the complete number of rows of the database d . This is because $f_2(d)$ only contains unique random numbers or plaintext encrypted with enc_p , permutations of the entries of each column cannot be detected after an application of f_2 and f_2 clearly is k -Ind-ICP.

We have shown that our privacy-aware folksonomy from Section 4 complies with the k -Ind-ICP security notion. Any folksonomy database secured this way is indistinguishable³ from a database with random entries of the same size.

6 Conclusion and Future Work

Folksonomies are popular to annotate and organize bookmarks, photos, or academic articles. However, folksonomies might also reveal personal information, e.g., interests, habits, behavior or social aspects. In this work we have proposed privacy-aware folksonomy systems. The main building blocks are as follows: We have distributed the folksonomy database among different providers. Each fragment of the database is encrypted. Our approach allows the users to control who is able to perform which operation on the folksonomy by sharing different keys with authorized persons. Our approach yields good performance for many prominent folksonomy operations. For operations that cannot be executed efficiently, we have described optimizations that do not disclose much additional information. Finally, we have evaluated our approach against a security notion. In particular, we have proven that each part of the distributed folksonomy is indistinguishable from a database containing random tuples.

As part of our future work, we plan to implement our approach in order to evaluate performance using real-world data sets from large-scale folksonomy systems.

References

1. Aggarwal, G., et al.: Two can keep a secret: A distributed architecture for secure database services. In: Proc. CIDR (2005)
2. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: CHI 2007: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2007)

³ Indistinguishability is a asymptotic property used in cryptography (c.f. [9]).

3. Bellare, M., Boldyreva, A., O'Neill, A.: Deterministic and efficiently searchable encryption. In: Menezes, A. (ed.) CRYPTO 2007. LNCS, vol. 4622, pp. 535–552. Springer, Heidelberg (2007)
4. Berkovits, S.: How to broadcast a secret. In: Davies, D.W. (ed.) EUROCRYPT 1991. LNCS, vol. 547, pp. 535–541. Springer, Heidelberg (1991)
5. Burghardt, T., et al.: Understanding user preferences and awareness: Privacy mechanisms in location-based services. In: CoopIS (2009)
6. Ceselli, A., et al.: Modeling and assessing inference exposure in encrypted databases. *ACM Trans. Inf. Syst. Secur.* 8(1) (2005)
7. Chaum, D., Crépeau, C., Damgard, I.: Multiparty unconditionally secure protocols. In: STOC 1988: Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing (1988)
8. Damiani, E., et al.: Balancing confidentiality and efficiency in untrusted relational dbms. In: CCS 2003: Proceedings of the 10th ACM Conference on Computer and Communications Security (2003)
9. Goldreich, O.: A note on computational indistinguishability. *Inf. Process. Lett.* 34(6), 277–281 (1990)
10. Goldreich, O., Micali, S., Wigderson, A.: How to play any mental game. In: STOC 1987: Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing (1987)
11. Hacıgümüş, H., et al.: Executing sql over encrypted data in the database-service-provider model. In: SIGMOD 2002: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data (2002)
12. Hacıgümüş, H., Iyer, B., Mehrotra, S.: Providing database as a service. In: Proceedings of 18th International Conference on Data Engineering (2002)
13. Heckner, M., Neubauer, T., Wolff, C.: Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. In: SSM 2008: Proceeding of the 2008 ACM Workshop on Search in Social Media, pp. 3–10 (2008)
14. Hore, B., Mehrotra, S., Tsudik, G.: A privacy-preserving index for range queries. In: VLDB 2004: Proceedings of the Thirtieth International Conference on Very Large Data Bases. VLDB Endowment (2004)
15. Khan, I., Schroeter, R., Hunter, J.: Implementing a secure annotation service. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 212–221. Springer, Heidelberg (2006)
16. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007 (2007)
17. Machanavajjhala, A., et al.: l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1) (March 2007)
18. Marlow, C., et al.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: HYPERTEXT 2006: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (2006)
19. Rivadeneira, A.W., et al.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: CHI 2007: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2007)
20. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10(5) (October 2002)

Seamless Web Editing for Curated Content

David Bainbridge and Brook J. Novak

Department of Computer Science
University of Waikato
Hamilton, New Zealand
{davidb,bjn8}@cs.waikato.ac.nz

Abstract. In this paper we present a new framework for editing that we have called Seaweed (short for *seamless web editing*) which enables authors to directly edit content on web pages within any common web browser—much like a word-processor—without the need of switching between modes. There are numerous ways to utilise the technique. This article reports on work integrating it with blogging software to support the direct creation and editing of curated content, and its subsequent evaluation through two field trials.

1 Introduction

The web has experienced several authoring paradigm shifts since its inception, and at each stage, the authoring process has been simplified. This article explores a new authoring process called *seamless web editing* (or Seaweed for short), which further simplifies things by being entirely modeless. This sets it apart from other methods, which enforce—we argue—an often unnatural distinction between viewing, editing, and publishing. Figure 1 shows an example of the technique in the context of editing metadata, in situ, in a digital library. The digital library system happens to be Greenstone [8], but in principle the approach could equally have been applied to one of the many other popular open source digital library systems: DSpace, ePrints, Fedora, etc.

The effect is like turning a web browser into a word processor without the need to reload the page—for any web page. (In this case, the page happens to be from a digital library.) This differentiates the technique from the approach used in systems such as GoogleDocs and the TinyMCE, where not only does the page need to be reloaded to activate editing, but the editing capability provided is reliant on this functionality being implemented natively in the web browser. Figure 1(a) shows the user browsing a list of titles. The last entry in this list contains one error and several undesirable artifacts—an enlarged version of this line can be seen in Figure 1(c). Having the year of publication embedded as part of the title is undesirable and deleting it would be an improvement. The capitalisation of the title is also inconsistent with the other titles presented (e.g., “For” with a capital letter). Finally, the title is also missing a word (it should be “for an” not just “for”).

Figure 1(b) shows the result after it has been edited using Seaweed—literally in a matter of seconds. Figure 1(d) shows the enlarged version. To delete the

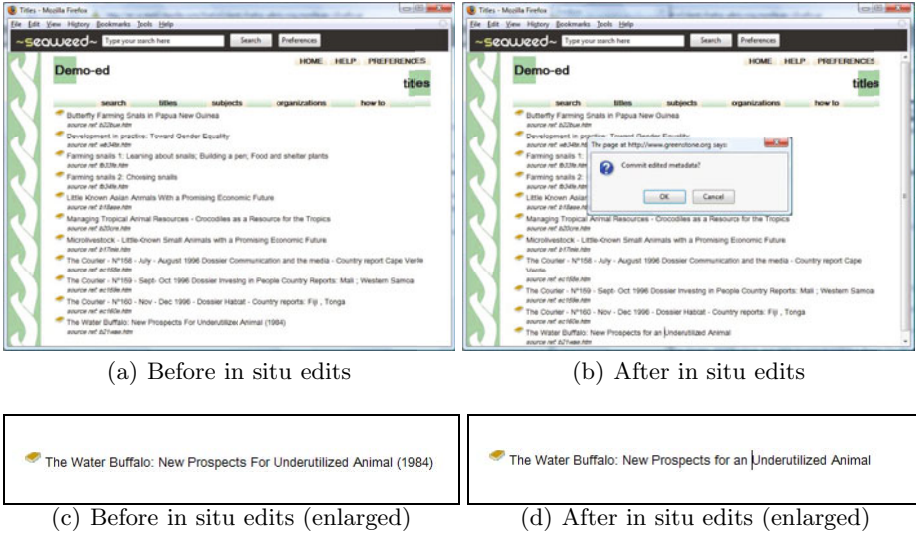


Fig. 1. In situ editing of metadata in a general purpose digital library

year, brackets around it and preceding space, the user simply selects this text with the mouse, and then presses delete. Clicking just after the “F” in “For”, a blinking cursor appears allowing them to delete the capital letter and then replace it with a lowercase version. Pressing the right arrow key three times they then type “an” followed by a space. All this is implemented to be as similar to the actions of a word-processor as possible, including copy, cut and paste, and a comprehensive undo facility. Satisfied with the edits, the user can then have them committed to the digital library—the popup window in Figure 1(b) gives the user the choice to commit or discard the edits.

2 Related Work

The closest form of web editing comparable with the technique described here is *in-place editing*, also called *live site editing* and *in-context editing* [4]. This is an editing model that allows users to edit content within the web pages they wish to change, however it is still a moded form of editing. In this section we review and contrast three pertinent examples: Sparrow, DirectEdit, and ISAWiki.

During the rise of Wikis and Blogs (before Web 2.0), a system called Sparrow was developed which hosts community driven websites [1]. Its key feature is a light-weight editing model, which as far as we know is the first in-place editing system for the web. Sparrow was designed to facilitate community shared web pages. It therefore shares similar community-based philosophies of a Wiki, but is distinct from a Wiki in that at the time of its development Wikis required contributors to have knowledge of HTML, whereas Sparrow did not. Furthermore, Sparrow had a finer level of granularity of editing: where users can edit parts

of a document as opposed to editing over a whole page as in a Wiki. Although Wikis have edit buttons at a section level of an article, users are directed to an editor containing the article's full content scrolled to the section that they want to edit.

DirectEdit is a content management system designed for small websites and small businesses. It features WYSIWYG in-place editing facilities to simplify the editing process [2]. In DirectEdit a document is broken down into sections which can be manipulated via a web browser. These sections are called DirectEdit elements for which there are five different types: Zones, Boxes, Fields, Images and Links. Fields are editable text areas which can contain formatting, using the same basic editing technique utilised by GoogleDocs and the like. Boxes are a combination of fields, images and links specified as an HTML template. The structure and formatting is defined by templates to adhere to the CSS design principle of separation of presentation and content/structure. Boxes can be used as reusable blocks containing a common design that is repeated in a single web page.

Pursuing the vision of Ted Nelson's Xanadu project of global editability, but in a web context, a system called ISAWiki was developed [6,7]. ISAWiki's design stemmed from both ISA (a desktop application for authoring that connected with a web-based server) and a hypermedia system called XanaWord [5]. The hand-crafted template system previously used in ISA became automated via a system called eISA, which used heuristic methods to identify document content and design elements (such as navigational menu items of a web page).

Two key inspirations drawn from the XanaWord project were: the ability to change content in a hypermedia system no matter who the original author might be; and the support and management of versioned documents. ISAWiki was designed to co-exist with the web: users would install a plugin for Internet Explorer 6 or Firefox (older versions now only supported) which would provide a sidebar. When a user requests a new web page, the plugin interrogates an ISAWiki server to check if personal modified versions created by the individual of the requested URL exist. The server returns a list of modified versions, and the plugin displays the latest version available. The user can view other versions via a list on the sidebar which is displayed while the user is in view-mode.

To edit and create a personal version of any page on the web, the user must click an edit button in the sidebar. The browser then enters an edit mode, where a WYSIWYG editing toolbar appears, and the document content (identified via the eISA subsystem) becomes editable. The in-place WYSIWYG editors maintain the exact CSS styles and layout. Once the user makes their change, they click a save button to save the new version.

3 Implementation

Seaweed is implemented as a client-side framework written entirely in JavaScript that exploits Document Object Model (DOM) manipulation to effect editing functionality. The net result is that all (or any part of) a web page can be

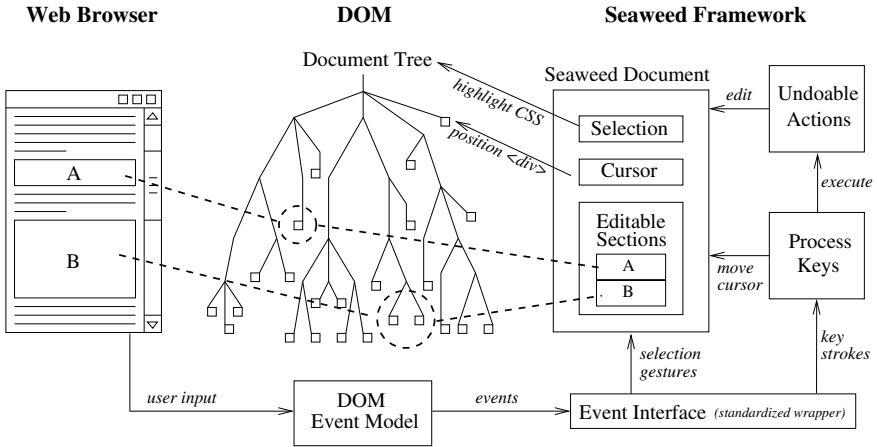


Fig. 2. Overview of the Seaweed framework in relation to the Document Object Model

made directly editable, without the need for the page to be reloaded (typically activating some form of browser-specific editing capability), as is done with the WYSIWYG editors reviewed in the previous section. Figure 2 shows an overview of the Seaweed framework and how it interacts with the DOM. The figure shows three views of a web page: the rendered display in the web browser seen and manipulated by the user (on the left), the underlying DOM tree which is manipulated via JavaScript (in the middle), and an abstract editable view seen by the Seaweed framework (on the right).

Similar to DirectEdit, in Seaweed CSS is used to separate document structure from content. In particular, a document is broken down into a set of editable sections (controlled by the standard *class* attribute of an HTML element being prefixed with the lexicon, “editable”). Figure 2 depicts these sections in the page labelled as “A” and “B”, also showing how they are part of the DOM tree and maintained by the Seaweed framework. Seaweed listens for all mouse and keyboard input events via an event interface: a sub-system that provides cross-browser event listening facilities. When the user clicks into editable content, a mouse DOM event is fired, normalised by the event interface and eventually, if needed, the Seaweed document model’s selection changes. When the user presses the “a” key, Seaweed interprets the key stroke and in effect will insert the letter “a” in the HTML document by changing the DOM—but only if the selection is within an editable section.

To support all this, fundamentally Seaweed needs to be able to determine, for a given mouse x, y position, which letter on the page it is closest to. Unfortunately there is no built-in provision in DOM to do this, however we have devised an algorithm to achieve this using a combination of JavaScript and on-the-fly insertion of *span* tags. Full details can be found in [3]. In brief, starting with the existing DOM operation that can determine which HTML block a mouse event occurred in, the technique dynamically inserts *span* tags to subdivide the

identified block, and then establish which of the two new blocks the event falls in. Iterating this process homes-in on the sought after character.

In essence the process is comparable to the game where one person tries to guess the number, say between 1 and 100, another person is thinking of. For each guess, the person is told if the sought after number is above or below that guess (or of course if they have struck the right number). The best approach on average is to first guess 50, and then, depending on the answer, subdivide the relevant number range in two and guess either 25 or 75 accordingly. For the character location algorithm, the 2D nature of the problem leads to some additional complications, however, these are all surmountable. Like the guessing game, the binary subdivision aspect to the algorithm leads to a $O(\log n)$ runtime complexity, where n is the number of characters in the identified initial block.

To give a practical sense to all this, we instrumented the character positioning algorithm and tested it over a wide range of web pages. In the case of clicking on the page to get the cursor to appear, the delay between clicking and cursor appearing was less than 200 ms, a delay which is essentially imperceptible to the user.

4 Creation and Editing of Curated Content

To evaluate seamless web editing we selected the example of creating and editing curated content with blogging software. To that end, we developed a suitable plugin for WordPress. In its most basic form the plugin allows the author (once they have logged in) to view their blog as others would see it *and* simultaneously have the ability to make any text edits instantly. The plugin provides much more than that, however. Figure 3 gives a glimpse of the full range of features. Notice the addition of dialogue windows to provide access to much of this expanded functionality, including control over forming new structural elements such as paragraphs, pages and posts. To help retain the immersive nature of the interface, dialogues are semi-transparent, and can be minimised as required.

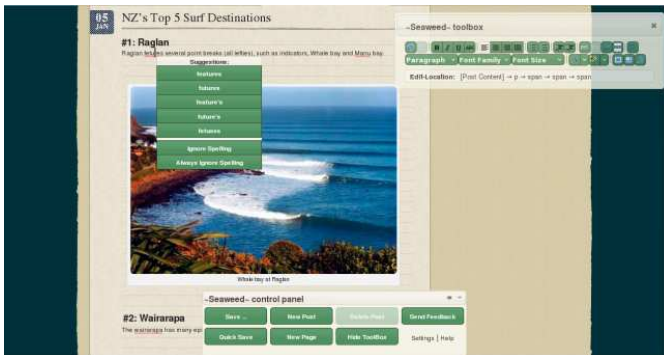


Fig. 3. The Seaweed WordPress plugin in use

There is a substantial array of plugins already developed for WordPress, and the Seaweed plugin is compatible with them, in the sense that it does not interfere with them. There are cases—such as the photo gallery plugin, that allows images to be added with a caption—that would greatly benefit from the seamless web editing approach; however, the nature of plugins in WordPress means there is no one central place the capability can be added that would mean it was automatically used by the other plugins. For evaluation purposes (see below) a modest set of widely used plugins were targeted and extended to use Seaweed: the afore mentioned photo gallery plugin being one of them, and can be seen in use in Figure 3.

5 Evaluation

Two user evaluations focusing on the Seaweed enhanced version of WordPress were conducted to establish its usefulness. In both cases logging of interactivity and pre- and post-questionnaires were collated and analysed. The first study took the form of a series of prescribed tasks, culminating in the participants being asked to blog about current events over a period of four days. They were instructed to create at least one post per day, and had the choice to create and edit their posts using either the standard editing facilities in WordPress or the Seaweed plugin. A total of nine participants took part in the prescribed study. All except for one participant had experience with blogging, using a range of blogging systems other than WordPress for at least one year. The accumulated logs totalled 205 entries.

The second evaluation was a study based on unprescribed tasks, “in the wild.” Participants who already had established blogs using WordPress switched to using the Seaweed enhanced version and continued to write their logs for a two week period, again choosing to use (or not use) the Seaweed features whenever they wished. Recruitment for volunteers for the second study was more wide spread, promoted primarily through postings on the WordPress community portal. From this a total of 26 participants installed the Seaweed plugin on their own blog, and registered to take part in the study. Of these 19 undertook sufficient interaction to be included in the activity log analysis. A total of 1009 log entries were captured for analysis.

Figures 4 and 5 provide a summary of the collated data. Figure 4 gives an overview of the size of edit operations performed. Edit calculations were based around a modified version of the Unix *diff* algorithm, where contiguous sections of text were classified as: inserted, deleted, replaced, or (trivially) unchanged. For non-trivial changes, the cost of each edit section was categorised as follows: minor 1–2; small 3–5; medium 6–10; and large 11+. Figure 5 summarises the questionnaire responses (TinyMCE is the open source visual editor that WordPress ships with).

Preference. Both the Likert responses and the activity logs of the prescribed study indicated that people without expertise using WordPress preferred editing content with the Seaweed plugin. The Likert responses in the unprescribed study

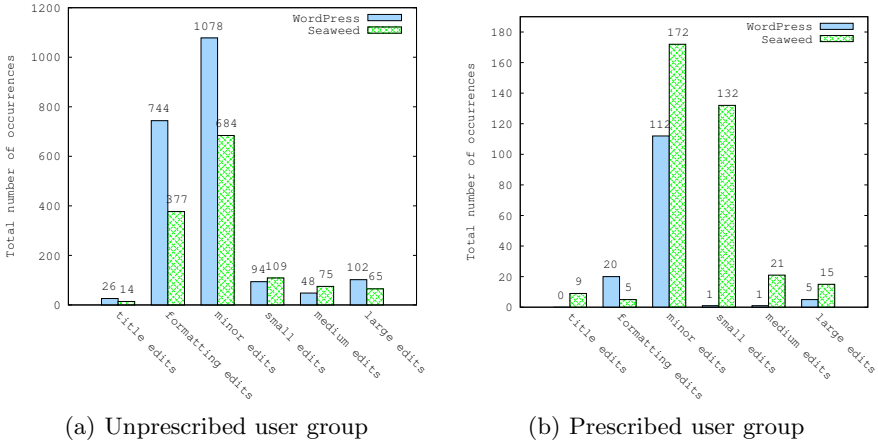


Fig. 4. Edit sizes based on analysis of logged interaction

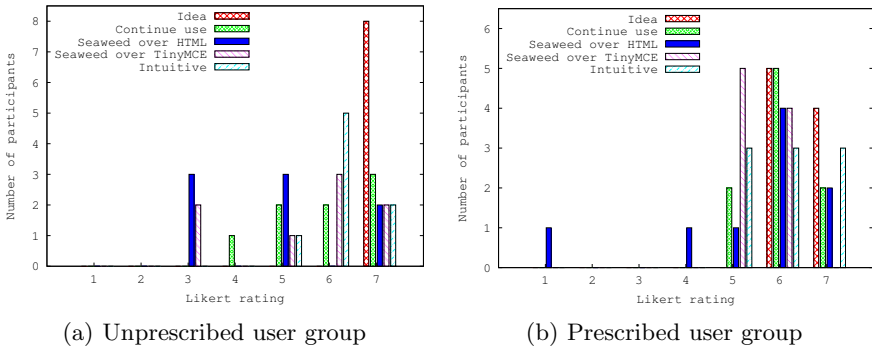


Fig. 5. Likert scale responses to post-questionnaire survey

likewise indicated that people with expertise using WordPress preferred editing content with the Seaweed plugin. Moreover, the full content analysis of the activity logs in the prescribed study revealed that participants clearly preferred Seaweed for making minor to medium sized edits. The analysis on the activity logs for the unprescribed study did not provide any distinct groupings.

According to the Likert responses in both of the studies, the participants generally preferred seamless editing over using HTML syntax directly. However it was found that there are a few users who always prefer editing HTML source over using visual editors in general.

Intuitiveness. In both of the studies participants had to teach themselves how to use the Seaweed plugin, thus participants were able to give authoritative feedback on the plugin’s intuitiveness. Overall, the Likert responses clearly indicated that the Seaweed plugin was highly intuitive. Generally, participants used the

Seaweed plugin in the unprescribed evaluation more than WordPress at the beginning of their observation period. The initial high usage activity indicates that users quickly adapted to Seaweed. Therefore, people who are accustomed to a moded way of editing can easily adapt to seamless editing.

6 Conclusion

To conclude, in this paper we have introduced the technique of seamless web editing and evaluated it in the context of authored personal content intended for public access. In the introduction we promoted Seaweed as modeless. This was to emphasise the key nature of the technique being described. It is of course a trivial matter to take a modeless system and develop one that enforces modes. Such developments with Seaweed would be natural for situations where an authoring environment wishes a strong distinction between editing and publishing. User authentication, another necessary ingredient, is already a capability of Seaweed. It was omitted from the examples to conserve space.

Our two field trials showed that the category of web page editing evaluated is characterised by a high number of minor edits—whether using the existing round-trip edit cycle provided or seamless editing. Comparing Seaweed with conventional editing, our analysis of the questionnaires showed a clear preference for the more direct form of manipulation.

References

1. Chang, B.-W.: In-place editing of web pages: Sparrow community-shared documents. In: Proc. 7th Int. WWW Conference, vol. 30, pp. 489–498 (1998)
2. Crosby, J.: What makes a CMS useful? Degree report. BCMS, Department of Computer Science, University of Waikato (2009)
3. Novak, B.J.: Seamless Web Editing. Msc, Department of Computer Science, University of Waikato (2010)
4. Robertson, J.: In-context versus back-end authoring. Step Two Designs (April 2008), http://www.steptwo.com.au/papers/cmb_incontext/ (Retrieved March 2010)
5. Vitali, F., Iorio, A.D.: A Xanalogical collaborative editing environment. In: Proc. 2nd Int. Workshop on Web Document Analysis, Liverpool (2003)
6. Vitali, F., Iorio, A.D.: Writing the web. *Journal of Digital Information* 5(1) (2004)
7. Vitali, F., Iorio, A.D.: From the writable web to global editability. In: Proc. 16th ACM Conference on Hypertext and Hypermedia, pp. 35–45. ACM, New York (2005)
8. Witten, I.H., Bainbridge, D., Nichols, D.: How to build a digital library, 2nd edn. Morgan Kaufmann, San Francisco (2010)

Automatic Classification of Social Tags

Christian Wartena

Novay, Brouwerijstraat 1, 7523 XC Enschede, The Netherlands
Christian.Wartena@novay.nl

Abstract. Collaborative tagging has become popular in recent years. As was noted in several studies completely different types of tags are found. Tags either can refer to the personal usage context of a tagger or can describe the tagged object. We investigate different types of tags found in LibraryThing, an online service in which books are tagged, and define a number of features that are typical for some of these classes. Finally, we show how these features can be used to classify tags automatically.

1 Introduction

The number of collections that is tagged by users is steadily growing. The set of tags assigned to an item by a large community of users will contain a lot of synonyms, spelling variants, alternative levels of details etc. that can enhance search but also can cause problems. A possible solution to these problems is to use statistically derived relations between tags ([1]).

Most work on tag similarities and tag clustering implicitly assumes that all tags are about the topic of the tagged item. Though this assumption is true for the majority of tags ([2]) there is also a non-negligible amount of tags that do not refer to the topic of the item, but e.g. to the media format of the item, the place the tagger stores a physical copy of the item, etc. In the present work we will investigate the possibilities to classify tags automatically in classes like "topic tags", "opinion tags" and "organizational tags".

There are various studies on automatic classification or clustering of tags into classes that are related to different topics. The work that is closest related to ours is [3], who classify tags for images into different categories like persons, artifacts, locations and groups. These classes are still rather different from the classes that we try to identify. Another difference is that we use inherent properties of the tags, while [3] exploit the possibilities of using Wikipedia.

The rest of this paper is organized as follows. First we discuss some classification schemata proposed for tags. In section [3] we define a number of features that can be used for classification, for which the results are given in section [4].

2 Tag Types

A number of classification schemata for tags have been proposed in literature. The majority of the differences is only in the degree of granularity. The classification schemata proposed by [4], [5], [6], [2] and [7] are given Table [1]. Most classes

Table 1. Tag classification schemes. Classes in the upper part of the table correspond to classes from other authors. The lower part gives the tags without correspondence.

Sen et al.	Xu et al.	Golder et al.	Bischoff et al.	Heymann et al.
Factual	Content-based	What it is about	Topic	Objective & content based
	Attribute	What it is	Type	
		Who owns it	Author/owner	Physical
Subjective	Subjective & Qualities	Characteristics	Opinions / Qualities	Opinion
Personal	Organizational	Task organization	Usage context	Personal
		Self reference	Self reference	
		Refining Categories	Time Location	Acronym Junk

can be aligned as suggested in the table. Some classes have no correspondence in the other schemata. In [4] refining tags are defined as tags that modify other tags and cannot be interpreted on their own. In most tagging systems, however, there is no possibility to modify tags by other tags, and relations like order of entering are not persistent. The categories *time* and *location* also do not fit very well in the table since they are in fact orthogonal to the other categories: location and time can refer to a topic, as well as to the usage context or can be used as self reference.

In the case study presented here we will concentrate on tags for books from LibraryThing. LibraryThing (www.librarything.org) is a web service for managing book collections, allowing, among other things, to tag books. As all tagged objects now are books, some tag classes get specific interpretations: Type now can be interpreted as genre and usage context is what usually is called reception.

It is also useful to have a look at other models used to describe traditionally archived objects. An interesting meta-data model especially suited for the library domain is described in the Functional Requirements for Bibliographic Records (FRBR, [8]). On the highest level this model distinguishes three groups of entities. The first group constitutes the core of the system and contains works, expressions, manifestations and items. The second group consists of persons and corporate bodies. The third group encompasses concepts, object events and places. Between entities relations can be defined. The most interesting type of relations are those between the entities of the first group: A work is defined as an abstract concept. An expression is the intellectual or artistic *realization* of a work, and thus still conceptual. Examples of expressions are translations or different editions. At the third level we find the manifestation as the physical embodiment of an expression. A manifestation is e.g. the printing of an edition. Finally, each manifestation has one or more individual items or exemplars. Each book in a library can now be described by attributes at all four levels and by its relations to other entities.

In the light of this model some problems of classifying tags mentioned above become much clearer. A tag in fact always is the attribute of some entity that can be either the described work itself (at one of the four conceptual levels), or another entity related to the work. Thus the classification in fact has to be two dimensional. For some tags a one dimensional classification indeed is very problematic, like for date and location that might be subject of a work, can refer to date and place of publication or to the acquisition of the exemplar.

While we find some tags related to the levels of expression and manifestation (e.g. *UK edition, translation*), the large majority of tags in LibraryThing refers either to the work or to the exemplar. At the level of works tags either describe entities that are in a subject relation to the work ('topics') or in a responsibility relation ('author'). All other tags at this level refer to attributes of the work. At the level of the exemplar we find tags like *borrowed* or *water damage*.

Considering the different classification schemata and the amounts of examples found for each class, we will use the classes as given in Table 2. Attribute includes genre but also other properties both at the level of work and expression, including the usage context. Self reference encompasses as well typical organizational tags as well as all tags related to the exemplar (usually owned by the tagger). Thus all tags expressing physical properties are also classified as self referential tags.

In many cases there are strong relations between topics authors and attributes. The genre historical fiction is closely related to the topic history, the author tag *Swedish author* of course relates to the attribute *Swedish literature* and so on. If we use the classification to improve topic based clustering of tags, search and recommendation it is probably enough to distinguish between *Self reference* and *opinion* at the one hand side and the other categories at the other.

3 Features of Tags

In the following we discuss a number of features that can be derived from the tagged data set itself, and that correspond to the classes defined above. Especially we are interested in distributional properties of tags.

Eccentricity. One of the most useful features turns out to be what we call the eccentricity of the tag. We base eccentricity on co-occurrence distributions of tags as defined in [9]. Tags that are highly related to one topic co-occur with a relatively small number of other tags. In contrast, we expect that tags that are not related to a specific object co-occur with many other tags.

Consider a collection of tagged items $\mathcal{C} = \{i_1, \dots, i_k\}$. Each tag occurrence is an instance of a tag t in $\mathcal{T} = \{t_1, \dots, t_l\}$ assigned by a user u in $\mathcal{U} = \{u_1, \dots, u_m\}$. Let $n(i, t, u)$ be the number of times a user u assigned tag t to item i , usually 0 or 1. To consider the tags assigned to an item we let $n(i, t) = \sum_u n(i, t, u)$ and similarly we define $n(u, t) = \sum_i n(i, t, u)$. Furthermore, let $n(t) = \sum_i n(i, t)$ be the number of occurrences of tag t , $N(i) = \sum_t n(i, t)$ the number of tags for i and $n = \sum_t n(t)$ be the total number of tag assignments. Now define

$q(t|i) = n(i, t)/N(i)$ the tag distribution of item i ,

$Q(i|z) = n(i, z)/n(z)$ the item distribution of tag z ,

$q(t) = n(t)/n$ the background tag distribution.

The probability distributions $q(t|i)$ and $Q(i|z)$ on the set of tags \mathcal{T} and the corpus \mathcal{C} , resp., describe how tag occurrences of an item i are distributed over different tags, respectively how the occurrences of a tag z is distributed over items. Now we can define the co-occurrence distribution of a tag z as: $\bar{p}_z^t(t) = \sum_i q(t|i)Q(i|z)$. The co-occurrence distribution is the weighted average of the tag distributions of items, where the weight is the relevance of d for z .

Now we define the eccentricity of a tag t as the Jensen Shannon divergence (see e.g. [10]) of the co-occurrence distribution of t and the background distribution q : $\text{eccentr}_\iota(t) = \text{JSD}(\bar{p}_t^\iota || q)$. We use the subscript ι to indicate that the co-occurrence distribution is computed using the co-occurrence on items.

As an example, consider the tags *must read* and *mysteries* in the LibraryThing dataset. The tags have similar frequencies (440 and 439 occurrences resp.) and occur on a similar number of items (378 and 383 items resp.). The first tag (*must read*) is clearly not about a topic while the second tag is. This is reflected by the divergence of the co-occurrence distribution with the general tag distribution, which is 0.0953 for *must read* and 0.223 for *mysteries*.

User Eccentricity. While a tag like *read 2003* co-occurs with arbitrary other tags when considering co-occurrence of tags on resources, we expect that such a tag is only used by a small fraction of all users, that also use tags like *read 2004* etc. In other words the co-occurrence distribution of these tags might diverge much more from the background distribution if we compute co-occurrence via user than via resources. Thus we define

$q(t|u) = n(u, t)/N(i)$ the tag distribution of item i ,

$Q(u|t) = n(u, t)/n(t)$ the item distribution of tag t .

The co-occurrence distribution of tags over users can now be defined as $\bar{p}_z^u(t) = \sum_u q(t|u)Q(u|z)$ and the user eccentricity of a tag as: $\text{eccentr}_\nu(t) = \text{JSD}(\bar{p}_t^\nu || q)$

Document Frequency. An important measure for term weighting in texts is the (inverse) document frequency. The document frequency of a tag t is defined as the number items (or documents) for which $n(t, i) > 0$. As usual we use the log of the document frequency.

Associated Ratings. In LibraryThing users do not only tag books but also have the possibility to rate them. Ratings are on a scale from 1 (half a star) to 10 (5 stars). Let $D_r \subset \mathcal{C} \times \mathcal{U}$ be the collection of pairs (d, u) such that the document d is rated by user u , and $D_r(t) \subset D_r$ be the subset of pairs such that the document d has been tagged with tag t . Define the average rating associated to a tag t as

$r(t) = \sum_{(u,d) \in D_r(t)} n(d,t,u)r(u,d)/n_r(t)$ where $n_r(t) = \sum_{(u,d) \in D_r(t)} n(d,t,u)$ is total number of ratings of items tagged with tag t . If $D_r(t) = \emptyset$, the average rating $r(t)$ is undefined. Note that if users rate a document at most once then $n(d,t,u) = 1$ for all $(d,t) \in D_r(t)$.

The associated average rating of most tags is close to the average of all ratings. Only associated ratings of tags expressing opinions are significantly higher or lower. For the simple linear classifiers we use the relative rating, defined as $r_{\text{rel}}(t) = |m - r(t)|$ where m is the average rating, gives slightly better results than the average ratings themselves.

Author Names. For all books in LibraryThing author names are available and displayed in user collections. Nevertheless, author names also appear frequently as tag. It is straightforward to check for each tag whether it is the name of an author in the dataset. In the experiment reported we have only checked for literal correspondence, and thus might have missed a number of variants. Some tags in our dataset with literal correspondence to author names have not been classified as author in the training data, like *dictionary* and *The Beatles*.

Common Substrings. Golder and Huberman [4] already mention that some words are characteristic for some classes of tags: the class self reference contains tags that start with *my*, like *my_favorite*. The category task organization contains a lot of tags starting with *to* like *to read*.

To find a number of useful substrings we first selected all words that are part of a tag (i.e. separated by blanks, dashes or underscores) and occur at least 10 times in our test set. From this set 8 words turned out to be useful for classification: *edition*, *author*, *reading*, *read*, *great*, *prize*, *award* and *favorite*.

4 Classifying LibraryThing Tags

We used a dataset from LibraryThing that was collected such that each user has supplied tags and ratings to at least 20 books and each book has received at least 5 tags ([11]). The dataset consists of 37.232 books tagged by 7.279 users with in total 10.559 different tags. The total number of tag assignments is 2.056.487.

4.1 Preprocessing

Before manual labeling of a training set and automatic classification we have merged a number of obvious variants of the same tag. This was done by selecting pairs of tags with similar spelling and distributions. For each tag, starting with the most frequent one, we select all tags with a small Levenshtein edit distance. We have assigned penalties to substitution, deletion and insertion that make variations at the end of the word cheaper than variation in the beginning and make changes of numbers more expensive than those of letters, that are in turn more expensive than changes regarding spaces, hyphens, underscores, etc. We select all variants with an edit distance-length rate that does not exceed a certain threshold. Since not all tags with a small edit distance are spelling variants

Table 2. Number of examples for each class

Class name (2 classes)	Class name (5 classes)	Nr. of examples
Work	Topic	150
	Author	100
	Attribute	150
User	Self reference	100
	Opinion	65

Table 3. Percentages of correctly classified instances using 10-fold cross validation

	5 classes	2 classes
baseline	27%	71%
all features	75%	92%
without substrings features	67%	89%

(consider e.g. the pairs *Hungarian literature - Nigerian literature* or *exploitation - exploration*), we require also that the distributions of the tags is similar. To measure distributional similarity we use the Jensen-Shanon divergence of the co-occurrence distributions that we require to be smaller than 0.2

The described procedure maps about 10% of the tags to another tag. Almost all variants that are found are real spelling variants or singular-plural pairs. For some longer tags also other variations are found like presence or absence of an article, abbreviations (*mt. everest - mount everest* or *Pulitzer winner - Pulitzer prize winner*) and word class variations (*transcendental - transcendentalism*, or *atheist - atheism*. Finally there is also a small number of real errors, like e.g. *16th century literature - 17th century literature*. While the number of errors is very low, the number of obvious candidates for matching that is not found is relatively large. We find a large number of synonyms that are not merged because the edit distance is too large. Most examples of this class are abbreviation (e.g. *ya - young adult literature*) and author names that either full names or initials. Very infrequent variants often are not detected since the the co-occurrence distributions are too different.

4.2 Manual Labeling

In order to test whether automatic classification into the classes discussed before is possible, we manually labeled 565 tags. First a random selection of tags was labeled. Subsequently, more examples of opinion and self reference tags have been sought in order to get more or less balanced classes. Also some additional examples of attributes were selected in order to include enough examples of the usage context and to include as well tags at the level of work and expression. This resulted in the number of examples for each class as given in Table 2.

4.3 Results

For classification we used the logistic linear classifier from the PR-tool box ([12]). All results were obtained by 10 fold cross validation. We evaluated for all 5 classes as well as for 2 classes. Since the 8 binary features based on the presence of certain substrings might be extremely dependent on the subset chosen and on the size of training data, we also trained the classifier without these features.

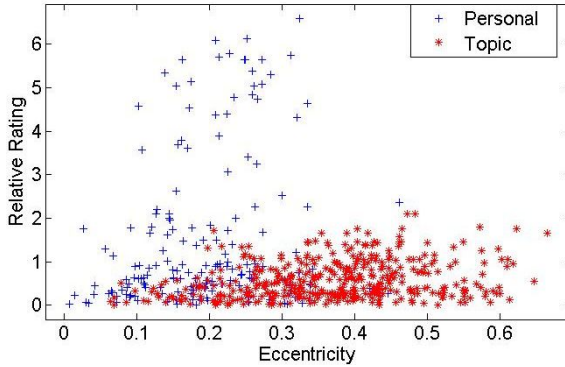


Fig. 1. Manual labeled data (two classes) in two dimensional feature space

Results are given in Table 3. As baseline we use the percentages of correctly classified tags when all tags are assigned to the most frequent class.

Closer inspection of the results shows that most severe errors stem from the substring features. E.g. almost all tags containing the word *author* are classified into the category Author. Counter examples like *Author I know* (self reference) are classified in the same way. Other types of errors usually are less severe, e.g. *didn't finish* was manually labeled as Self reference, but classified by the algorithm as Opinion. The features with the largest contribution are rating and eccentricity. Fig. 1 shows how these features can distinguish between personal and work-oriented tags.

We also classified the complete dataset of 9501 tags with a classifier trained on all manually labeled examples. In the case of two classes 722 tags were labeled as personal. Fast inspection of the results suggests that about 150 tags were misclassified. After correction 637 tags were classified as personal tags, suggesting that about 6 to 7 % of the tags in the collection belongs to this category. This result is in line with the manual classification of 2000 LibraryThing tags by [7], who found that 1.8% of the tags expresses an opinion and 6.15% of the tags is personal or related to the owner.

5 Conclusion

Tagging has become an important feature of many Internet based collections. The lack of any structure contributes to the ease of tagging and thus probably to its popularity. However, this also has as a consequence that different types of tags, like descriptive tags, tags expressing opinions or personal tasks, are mixed up. In the present paper we have investigated the possibility of classifying tags into different types. In order to classify tags automatically into the proposed categories we have defined a number of features derived from the tagged collection. Features are based on the characteristics of tag distributions, on ratings given in combination with tags and on common substrings of tags. Using these features it was possible to classify the majority of tags correctly.

For applications that suggest classifications to users, e.g. to help them organize their tags, the degree of accuracy reached might already be acceptable. In order to improve the quality of the classification other features have to be investigated, especially features relying on other resources like dictionaries (e.g. Wordnet) or encyclopedia. Another topic for future research is the effect of classification on retrieval or recommendation. It can be expected that classification of tags into different types offers important possibilities to improve many tag-based tasks and systems.

Acknowledgments

The research presented in this paper was conducted during a visit of the author at the Technical University Delft, supported by the PetaMedia Network of Excellence, funded by the European Community's Seventh Framework Program under grant agreement n° 216444. Part of the research was done within the MyMedia project (FP7 grant agreement n° 215006).

References

1. Wartena, C., Brussee, R., Wibbels, M.: Using tag co-occurrence for recommendation. In: ISDA, pp. 273–278. IEEE Computer Society, Los Alamitos (2009)
2. Bischoff, K., Firan, C.S., Nejdil, W., Paiu, R.: Can all tags be used for search? In: CIKM. pp. 193–202 (2008)
3. Overell, S.E., Sigurbjörnsson, B., van Zwol, R.: Classifying tags using open content resources. In: Baeza-Yates, R.A., Boldi, P., Ribeiro-Neto, B.A., Cambazoglu, B.B. (eds.) WSDM, pp. 64–73. ACM, New York (2009)
4. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. *J. Information Science* 32(2), 198–208 (2006)
5. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: Collaborative tag suggestions. In: Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland, Citeseer (2006)
6. Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: Hinds, P.J., Martin, D. (eds.) CSCW, pp. 181–190. ACM, New York (2006)
7. Heymann, P., Paepcke, A., Garcia-Molina, H.: Tagging human knowledge. In: Davison, B.D., Suel, T., Craswell, N., Liu, B. (eds.) WSDM, pp. 51–60. ACM, New York (2010)
8. IFLA Workgroup on Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records: Final Report. K.G. Sauer, München (1998)
9. Wartena, C., Brussee, R.: Instance-based mapping between thesauri and folksonomies. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 356–370. Springer, Heidelberg (2008)
10. Cover, T., Thomas, J.: Elements of information theory. Wiley Interscience, Hoboken (2006)
11. <http://dmirlab.tudelft.nl/users/maarten-clements>
12. van der Heijden, F., Duin, R.P., de Ridder, D., Tax, D.M.: Classification, parameter estimation and state estimation - an engineering approach using Matlab. John Wiley & Sons, Chichester (2004)

Exploring the Impact of Search Interface Features on Search Tasks

Abdigani Diriyé¹, Ann Blandford¹, and Anastasios Tombros²

¹ University College London Interaction Centre, University College London, UK

² Department of Computer Science, Queen Mary University London, UK
{a.diriye,a.blandford}@ucl.ac.uk, tassos@dcs.qmul.ac.uk

Abstract. There is growing recognition that exploratory search is less well supported by existing search interfaces than known-item search. In this paper, we report on a study in which three interfaces providing different levels of search support were developed and tested, for both known item and exploratory search tasks. A rich qualitative analysis of participants' search behaviours and perceptions was conducted. As expected, the simplest interface provided better support for known item than for exploratory search tasks. Conversely, richer search interface features were found to provide better support for exploratory search, but would distract people from the objective of more clearly defined search tasks. This study provides preliminary evidence that searching is most effective when supported by an interface that is tailored towards the search activities of the task.

1 Introduction

People need to find information to support the many kinds of information activities they undertake – from knowing what time their train leaves to developing a rich understanding of a new topic. In order to do this, they need to locate information that is appropriate to their current state of knowledge, and that helps them move towards understanding [1]. In order to look for information within the Digital Library (DL) or web resource they have at their disposal, people need to be able to frame their information needs in terms that are appropriate to the available domain and knowledge resources [6]. Hence, there is a need to support query formulation so terms are produced that will retrieve suitable documents for information use. In this paper we look at query term suggestions and try and understand how richer means of support affect the different kinds of search tasks people undertake.

2 Background

People's day-to-day search activities can vary greatly in their motivations, objectives, and outcomes. The body of literature has classified search tasks into two over-arching categories: known-item and exploratory search tasks (some researchers have referred to these categories as simple and complex, or closed- and

open-ended) [7, 8]. Known-item search tasks usually involve looking up some discrete, well-structured information object: for example numbers, names and facts [8]. Exploratory search tasks, on the other hand, are seen to be more complex and involve investigating, learning and synthesis of information [7, 8]. What really differentiates known-item and exploratory search tasks is the clarity of the information need, the familiarity the searcher has with the task domain, and the analysis and understanding involved [15]. These factors invariably affect how searchers interact with information, and how they search and browse.

To support people's various search tasks, a number of search interface features have been developed. One interface feature that has become widely adopted in DLs is Interactive Query Expansion (IQE). IQE helps the searcher to expand their query with system-generated suggested terms. However, the suggested terms provided by the system can lack meaning and context when the searcher is unfamiliar with the domain or vocabulary. One novel method to supplement suggested terms is with context [5]. Such richer support is clearly needed to further improve existing search features such as IQE, but it is not clear in the literature what impact richer support has.

To be able to construct more usable DLs, an understanding of how search interface features such as query term suggestions impact people's information-seeking is needed, to ensure we design interfaces that truly support people's information-seeking. This is the motivating factor behind this study. The study was exploratory and observational in nature, and only sought to document and analyse interesting phenomena. Like [11], the data presented in this paper is illustrative of the general search behaviours observed.

In the remainder of this paper, we detail the design of our system and study, and discuss the implications our results have for search interface design.

3 System Design

The experimental system we used in this study had three different interfaces, each providing incrementally richer search support. Each interface assisted in searching, browsing and understanding of information in different ways, and was constructed using existing web technologies (i.e. Yahoo! and Google API).

The baseline interface (baseline Search Friend) represents the lowest level of search support, and only facilitates searching and document selection. The baseline interface (Fig. 1a) resembles the layout and functionality of popular search interfaces such as Yahoo! and Google.

The intermediate interface (Search Friend I) is the next level up from the baseline in terms of functionality, and provides suggested query terms along with the search results. The suggested query terms can be used to assist in query reformulation and to also filter the result set so documents containing the suggested query terms are displayed (Fig. 1b). An informal evaluation comparing our custom query expansion algorithm against suggested query terms from Ask, Google and Yahoo! identified Yahoo! as the best source for highly relevant and meaningful suggested query terms.



Fig. 1. a) Baseline interface of Search Friend; b) Snippet of additional feature in Search Friend I; c) Snippet of additional feature in Search Friend II

The full interface (Search Friend II) provides the richest level of system support and functionality. Along with providing a set of search results, it also offers suggested query terms and the context in which the suggested query term exists. The context of the suggested query term was generated by extracting the highest scoring sentence that contains the suggested query term; this representation was used to supplement the suggested query term because of the advantage a dynamic summary has in conveying document relevance over a more static definition [12]. The full interface presents information in a manner that strives to elucidate the concepts in the suggested query terms.

4 Study Design

The study was a 3x3 laboratory-based within-subject experiment. There were 3 different levels of system support and 3 different kinds of search tasks. We used the Search Friend system to investigate the role richer search interfaces play during different search tasks. Each participant was expected to carry out a search task on each one of Search Friend’s interfaces systematically. There were eighteen participants in total, and they were recruited in person. All had minimal knowledge of the topics of the search tasks, and a minimum of 6 years computing

Table 1. Search task types and narratives

Task Type	Task Name	Task Statement
Simple Known-item Search	Human Smuggling	Identify incidents of human smuggling
Complex Known-item Search	Wrongful Convictions	Find documents that discuss freed prisoners who have been wrongfully convicted based on faulty forensic evidence, poor police work, or false testimony
Exploratory Search	Racial Profiling	How have instances of racial profiling encroached upon the civil liberties of individuals, and has legislation changed as a result?

experience. Their ages ranged from 22 – 41 years, and they comprised 12 males and 6 females. They all rated their computing proficiency between average to excellent, and all frequently conducted online searches. The search tasks they were asked to carry out were: a simple and complex known-item search tasks, and an exploratory search task. This was so we could examine the effects across different search tasks. We employed qualitative and quantitative data gathering methods such as think-aloud protocols, screen-recording of interface interactions and questionnaires to be able to investigate the interplay between the interface features, the user and the search task.

4.1 Search Tasks

As we are investigating the impact richer search interfaces have, a spectrum of search tasks covering different search task types and goals would ideally need to be used. Given the obvious constraints, a trade-off had to be made between getting a broad representative sample of search tasks and what was feasible. To this extent, we have focussed our attention on a small subset of search tasks from the two overarching search task categories. For the known-item search tasks, searchers were required to identify a known piece of information, and for the exploratory search task, the objective was to address some general topic or open question. These search tasks were obtained from the TREC tracks, and their search task categories were determined based on the search task’s objective, complexity and difficulty; Table 1 describes the search tasks in detail.

4.2 Study Procedure

To nullify the effects of learning and fatigue, a Latin square design was used to permute the search tasks and system interfaces. Each user study lasted up to an hour, and began with the participants being asked to fill out a consent form and provide answers to a short questionnaire that elicited demographic information. For each interface, a demonstration was given, and a written statement of

the search task was provided to the participant. The participants noted down on paper their understanding of the search topic before and after the search task, and were given at most 10 minutes to complete it. Think-aloud protocol was collected from the participants, and their interactions with the system were recorded using screen-capturing software. After the search tasks, an exit questionnaire was administered to elicit the participants' disposition towards the system interfaces and their general experience.

4.3 Data Gathering and Analysis Methodology

Once the data had been gathered, participants' think-aloud and screen-recording data was documented by transcribing the think-aloud data, and then overlaying the screen-recording data onto the think-aloud transcriptions. This approach allows us to collate rich textual data from different but complementary data sources to understand and analyse "*ground truths*" of the search behaviours. Emergent themes and patterns were identified in the data and then assigned codes to the different categories and concepts that had arisen from the data. The coding scheme was constructed in an iterative manner that entailed going through the transcripts and repeatedly refining the codes. The unit of analysis in this study varied from a single spoken sentence during the think-aloud, to several instances of user-system interactions, this was necessary because it was not always possible to identify concepts from just single sentences. In this paper we mainly focus on the qualitative aspects of the study to understand the underlying causes of the phenomenon we were investigating. A qualitative approach would provide the richness in the data that would reveal the complex processes at play, and ensure we can understand the data in its context. This kind of approach would allow us to go beyond the "*How fast?*" and "*How many clicks?*" to an understanding of why things truly happen.

5 Results

In this section, aspects of the study relating to search behaviours, interface features, and participant perceptions are presented.

5.1 Search Tasks

To validate our expectation that the different search tasks would be perceived differently, we gathered data using semantic differentials measuring how complex, difficult and vague the participants thought the search tasks were. A statistically significant difference (i.e. one-way ANOVA) was found at $p < 0.01$, and as we can see in Fig. 2 the simple known-item task was perceived as having the least vague information need ($F(2,53)=5.08$, $p=0.0097$), and being the least complex ($F(2,53)=12.69$, $p < 0.001$) and difficult ($F(2,53)=5.40$, $p=0.0074$) search task, compared to the complex known-item and exploratory task. Tukey post-hoc tests identified significant differences lay only between the simple known-item, and the complex and exploratory tasks ($p < 0.05$, for all).

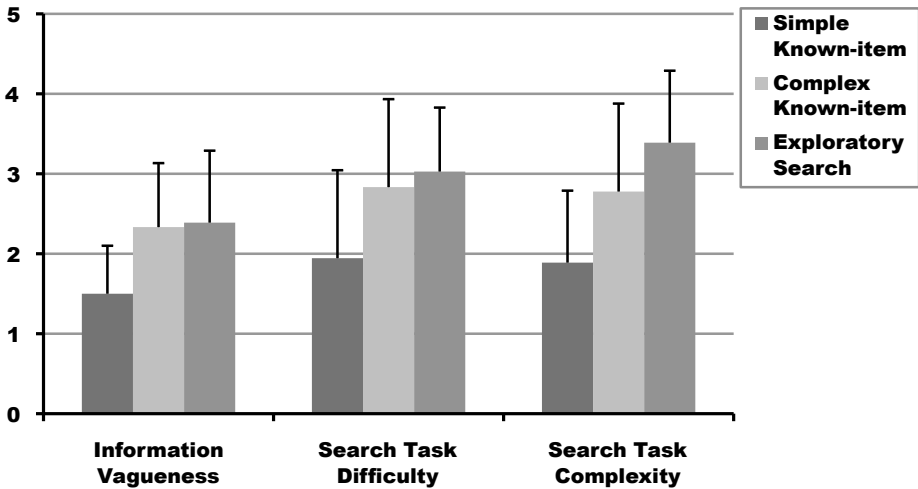


Fig. 2. Participants' views of search tasks

5.2 Information-Search Behaviours

To understand the effects richer search features have in the context of search tasks, a task-centric perspective is taken where the expected actions and goals of the search tasks are compared with the actual search behaviours observed.

Simple Known-Item Task. The simple known-item task required identifying incidences of human smuggling. The task was well-defined, hence the participants had a clear and static concept of the information they needed to find. The participants would therefore need to address this search task by navigating between documents, and locating specific items of information.

On the baseline interface, the participants' behaviours comprised rudimentary search tactics and strategies such as query formulation, document examination, selection and analysis. The participants' relevance judgements were focussed on identifying instances of human smuggling, and documents were selected that could be used to do this. The simple design of the interface facilitated the participants to carry out these simple search behaviours.

The Search Friend I interface was used by the participants in this search task to filter the search results and identify document topics and suggested query terms. During the search sessions no instances were observed where the participants used the interface to carry out investigation, examination or analysis of the information. Instead, the interface features complemented the participants' search behaviours by allowing them to refine the search result set so documents containing specific topics appeared, or assisted in identifying possible suggested query terms.

On the Search Friend II interface, similar behaviours were observed to the baseline and Search Friend I interface like: look-up, scanning, filtering, and document selection. Instances were observed where the richer level of support (i.e. the contexts of the suggested terms) enabled the participants to explore and investigate other related concepts and topics. The presence of richer search support spurred the participants to construct an understanding of the related concepts on the interface; in the context of the search task, this did not enable the participants to directly address the search task as this led participants to explore tangentially related concepts (2/8 participants were observed to do so), as the excerpt from participant 16 suggests:

“So I’m just going to click reset to see what the overall terms are”
 <User clicks ‘reset’ button
 “Snakehead, here’s a term I’ve never heard before, so I’m going to click on it”
 <User selects document –P16

Complex Known-Item Task. Like the simple known-item task, this search task was well-defined so obvious similarities were present in participants’ search behaviours. We observed a tendency by the participants to frequently navigate between documents, scan, search, identify and verify information. But, unlike the simple known-item search task, there was an inherently higher level of complexity because of the number of criteria to be fulfilled to address this search task.

On the baseline interface, the participants were observed to address this task predominantly by scanning and browsing for content that addressed aspects of this search task. In contrast to the simple known-item task, because of the demands of the search task, there was a tendency for the participants to be analytical and interact with the information more (15/18 of all participants observed):

<User selects document
 <User tries to locate term “political” within document and uses search function
 “OK this one seems to be specifically about political prisoners. I’m looking for a particular case, it seems to be generally statistical, but it has a back issue of...” – P2

The Search Friend I interface enabled the participants to locate relevant documents by filtering and using terms they recognised as relevant to identify documents relating to the search topics. During this search task, instances were not observed where the absence of contextual information adversely affected their searching. This is down to the task involving essentially the look-up of information and not more complex activities such as learning, comparison or investigation.

The richness of the Search Friend II interface was better utilised by the participants during this search task. It was used to filter and identify relevant documents and concepts, and preview selected snippets on documents the participants felt were relevant. On occasions when participants were uncertain of a concept

or of a specific suggested query terms, this richer level of interface support provided clarification, and possible avenues for further searching. Exploration and investigating incidental information did not necessarily address the search task directly, but it did provide a richer understanding of the topic, as can be seen from participant 10:

<User clicks on suggested query term

"There is something called the innocence project and I don't know what that is."

<User clicks on suggested query term 'innocence project'

"So this [document] is a website about wrongful convictions, as I mouse over it the suggested terms are... well it looks like some project set up to look up wrongful convictions"

[User reads suggested query term's context]

"OK we have to look at this document" – P10

The participants' search behaviours indicate that when they had a clear idea of their information need, and were faced with a complex and demanding search task, the level of support provided on both of the Search Friend systems could be successfully used to address their task.

Exploratory Task. Because of the nature of the exploratory search task, substantially more analysis and comprehension of the information was observed during this search task. To address this search task, the participants were observed attempting to try and collate information and construct an understanding of the topic.

On the baseline interface, a large portion of the participants' behaviours comprised document navigation; where the interface would be used to identify relevant documents to navigate to. The absence of interface features that encouraged interaction with the information meant that on the baseline interface, the participants did not engage in much analysis and comprehension of topics, concepts and document surrogates whilst on the search results page. This was observed to happen almost exclusively within the documents where the participants would have to manually seek these topics and concepts within the document to make sense of them.

Like the simple and complex search tasks, the Search Friend I interface was widely adopted by the participants to filter, assist in relevance judgement of the documents, and identify relevant suggested query terms. This interface was more than adequate when the participants were engaged in looking-up documents and topics, but a lack of support was evident when participants engaged in investigating, analysing and understanding the information. When the participants were trying to understand what topics to explore and how the information related, the absence of contextual information for the suggested query terms resulted in the participants being uncertain about the relevance of the term, and its relation to the search task (4/13 participants were observed to have some difficulty). This led some searchers to be uncertain of which concepts to explore; participant 14 illustrates this particularly well:

<User enters 'racial profiling'
 "So there is a definition of racial profiling"
 <User browses document's suggested terms
 "I don't understand what [the suggested term] traffic stops has to do with racial profiling, African American, OK, police officers – maybe they're involved in minorities – yeah, maybe they're involved in..." – P14

Finally, for the Search Friend II interface, the participants' tackled this search task in a similar fashion to the other two interfaces by exploring relevant topics and formulating an understanding of the information. But, the interface features on Search Friend II enriched the participants' searching and browsing by providing possible search topics to investigate, as well as clarification of the relationships between the concepts and relevant topics in the search result set. The additional information and context provided by the Search Friend II interface allowed the participants to formulate a richer understanding of the relevance of the suggested query terms and their relation to the search results. This gave the participants freedom to elaborate on a particular document, or suggested query term when they felt uncertain of its relation to their information need:

"There's a thing called traffic stops, all the others are obvious"
 <User clicks on suggested query term "traffic stops"
 "And I'm looking at what it says about traffic stops"
 <User hovers over search result
 <User clicks on suggested query term "traffic stops"
 "And I'm just getting a definition of traffic stops because it's not something I'm familiar with." – P2

5.3 The Role of Interface Search Features

Further examination of the screen-recording and think-aloud data identified three possible ways interface features such as suggested query terms impact participants' information seeking. Interface features can facilitate, transform or impede participants' information-seeking. The examples described in Table 2 are illustrative of these behaviours.

Facilitate. Interface features can facilitate search actions that help in completing a search task. For example, participant 8 (c.f. Table 2) is trying to browse documents to locate relevant information. The interface features support querying and browsing which are used to address an aspect of the search task. In such cases where the interface features provide support for the user's task, the search interface feature can be considered transparent as the focus of the user's attention is on the search task.

Impede. Conversely, the interaction between the user and system can result in their attention being centred on the interface feature and not on the search task. This happens when the interface feature does not provide search actions conducive to completing the search task or is non-obvious. In this instance, the search interface feature is opaque, and completion of the search task is subverted.

Table 2. Effects of the interface features on participants' information-seeking

Code	Definition	Example
Facilitate	Supporting the user to achieve some search action.	“OK I'm gonna start by searching for some key terms such as freed prisoners.” <User enters “Freed prisoners” “OK, I'll just read through the list” <User browses search results – P8
Impede	Search support that hampers or hinders information-seeking.	<User browses search results “The concepts on the left though, don't seem to be helping me out too much. It just got stuff like police department – it's annoying because there not enough information anywhere to tell me what the links are before I click on them” – P3
Transform	Provide alternative search behaviours that support the user in their information-seeking.	<User enters “US racial profiling” as query “I'm actually hovering. The first thing I did there without thinking about it is hover over the search terms and just read down the left hand column and see what was coming up.” <User browses search results – P3

Transform. The use of interface features can also transform and provide alternative search strategies to the user. This can be the transition from one mode of searching such as successive querying and browsing to filtering. In Table 2, the interface feature transforms participant 3's search behaviour by providing a novel way of browsing which they utilise. The interface features transform information-seeking when some search-task related strategy is being supported.

6 Discussion

In line with previous work [3, 7, 8, 15], our results have shown that different categories of search tasks are understood and perceived differently, this in turn has a knock-on effect on information-seeking behaviours. We have seen that known-item search tasks are characterised more by focused and direct search behaviours where the searchers have a clear understanding of their information need, whereas for the more complex and exploratory tasks, the vagueness in their information need results in more exploratory browsing and searching. The findings that have transpired from this study also suggest a relationship between search tasks and search interface features.

Known-item tasks comprising rudimentary search actions such as look-up and verification were effectively supported on the baseline and Search Friend I interfaces. But for the more exploratory search task, which comprised higher-level search activities such as analysis, investigation and comprehension, we have seen participants have difficulties making sense of the information, and an absence

of support for these search activities on these interfaces. The exploratory search task was better supported on Search Friend II as it enriched the participants searching and browsing by providing richer information and context. The participants were able to formulate a richer understanding of the relevance of the suggested query terms and their relation to the search results. This supported search activities like analysis, investigation and comprehension better. Also, as a result of the rich support on the Search Friend II interface, these higher-level search activities were also exhibited on the known-item search tasks.

Ultimately, interaction with search interface features can transform and facilitate search actions that enable search tasks to be addressed. Conversely, we have also seen that as well as supporting and transforming a searcher's information-seeking, it can also impede and distract the searcher from their search task. This is prevalent when the search actions being provided are not obvious, or are not supportive of search actions integral to accomplishing the search task.

Vakkari [13], Belkin [2], Payne et al. [10] and Wilson et al. [16] have all discussed the correlation between interface features and the users' goals. They suggest that a mapping exists between the actions a system provides and the goals searchers try and achieve. Particularly, Payne et al. have noted that some of the system features do not directly address the user's goal, but provide shortcuts and more efficient performance of the task. But, Carroll has noted that as well as allowing the user to efficiently address their search goal, system features can also hinder movement towards their goal [4].

7 Conclusion

In this paper we have presented an exploratory study that investigated the impact richer search support has on information-seeking across known-item and exploratory tasks. This study provides preliminary evidence that searching is most effective when search interface features facilitate search activities that are associated with the task.

We have seen that search interfaces affect people's information-seeking by transforming and facilitating search actions or by impeding and distracting them from the focus of the search task. The baseline Search Friend interface provided better support for simple known-item tasks, compared to the more exploratory search task, because search activities integral to the task were not supported. Conversely, the richer Search Friend II interface provided better support for the complex and exploratory task, but distracted people from the objective of the more clearly defined simple search task.

As a qualitative study, this work has focused on one set of interfaces and one set of tasks. The approach, and the findings appear to be promising, and they suggest avenues for further research to test how well these findings generalise to other forms of search support and for a wider range of tasks (both known-item and exploratory). This work represents one "*data point*" towards understanding exploratory search behaviours and how to support such behaviours through design.

Acknowledgements

A special thank you to Sarah Faisal, Stephann Makri, Max Wilson, Simon Atfield and our anonymous reviewers for their constructive comments. This work is supported by an EPSRC studentship.

References

- [1] Belkin, N.J., Oddy, R.N., Brooks, H.M.: Ask for information retrieval: part I: background and theory. In: *Readings in information Retrieval*, pp. 299–304. Morgan Kaufmann Publishers, San Francisco (1982)
- [2] Belkin, N.J.: Interaction with texts: Information retrieval as information-seeking behavior. In: *Information retrieval 1993 Von der Modellierung zur Anwendung*, pp. 55–66. Universitaetsverlag Konstanz, Konstanz (1993)
- [3] Bystrom, K., Jarvelin, K.: Task complexity affects information seeking and use. *Inf. Process. Manage.* 31(2), 191–213 (1995)
- [4] Carroll, J.M.: *HCI models, theories, and frameworks: Toward a multidisciplinary science*, 1st edn. Morgan Kaufmann, San Francisco (April 2003)
- [5] Diriyeh, A., Blandford, A., Tombros, A.: A polyrepresentational approach to interactive query expansion. In: *Proc. of JCDL 2009*, pp. 217–220 (2009)
- [6] Fields, B., Keith, S., Blandford, A.: Designing for expert information finding strategies. In: *People and Computers XVIII*, pp. 89–102 (2005)
- [7] Li, Y., Belkin, N.J.: A faceted approach to conceptualizing tasks in information seeking. *Inf. Process. Manage.* 44(6), 1822–1837 (2008)
- [8] Marchionini, G.: Exploratory search: From finding to understanding. *CACM* 49(4), 41–46 (2006)
- [9] Qu, Y.: A sensemaking-supporting information gathering system. In: *Proc. of SIGCHI Conf. CHI 2009*, pp. 906–907 (2003)
- [10] Payne, S.J., Squibb, H.R., Howes, A.: The nature of device models: the yoked state space hypothesis and some experiments with text editors. *Hum. Comput. Interact.* 5(4), 415–444 (1990)
- [11] Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R.: The perfect search engine is not enough: a study of orienteering behavior in directed search. In: *Proc. of the SIGCHI Conf. CHI 2004*, pp. 415–422 (2004)
- [12] Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: *Proc. of 21st ACM SIGIR Conf.*, pp. 2–10 (1998)
- [13] Vakkari, P.: Task-based information searching. *Annual Review of Information Science and Technology* 37, 413–464 (2003)
- [14] White, R., Marchionini, G.: Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.* 43(3), 685–704 (2007)
- [15] White, R., Roth, R.: *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Series on Information Concepts, Retrieval, and Services (2009)
- [16] Wilson, M.L., Schraefel, M.C., White, R.: Evaluating advanced search interfaces using established information-seeking models. *J. Am. Soc. Inf. Sci. Technol.* 60(6), 1407–1422 (2009)

Relevance in Technicolor

Ulises Cerviño Beresi¹, Yunhyong Kim¹, Dawei Song¹,
Ian Ruthven², and Mark Baillie²

¹ The Robert Gordon University, School of Computing

² The Strathclyde University, Department of Computer and Information Sciences

Abstract. In this article we propose the concept of relevance criteria profiles, which provide a global view of user behaviour in judging the relevance of retrieved information. We further propose a plotting technique which provides a session based overview of the relevance judgement processes interlaced with interactions that allow the researcher to visualise and quickly detect emerging patterns in both interactions and relevance criteria usage. We discuss by example, using data from a user study conducted between the months of January and August of 2008, how these tools support the better understanding of task based user valuation of documents that is likely to lead to recommendations for improving end-user services in digital libraries.

1 Introduction

Faced with the decision of whether or not to retain a piece of information in their personal collection, individuals engage in gauging the value of a document. This is distinct from a binary judgement regarding whether the document is relevant or not relevant. The situation is akin to the valuation process used by an antique dealer in assessing the value of a artefact: several criteria are employed to determine the object value, e.g. in terms of date, rarity, popularity, and condition of the object. Likewise, the qualitative or pragmatic value of a document is determined by a number of criteria, e.g. currency, novelty, validity and clarity. The consideration of these criteria results in an overall estimate of the document's usefulness within the context of user tasks. The criteria employed, although clearly related to metadata elements employed within libraries (e.g. Dublin Core Metadata Elements <http://dublincore.org/documents/dces/>), as well as the topicality of the document, do not map directly onto either of these. By studying the way in which information searchers and seekers utilise and weight these criteria, we hope to bridge the gap between human information valuation behaviour, and implementations of information retrieval (IR) engines and library end-user services.

To be able to study these criteria one must observe them first, and do so in a realistic scenario. The guidelines for evaluating IR systems proposed by Borlund [4] allows the researcher to gather both performance as well as cognitive data; data which includes these relevance criteria observations. Realism is achieved as the framework involves potential end-users as test persons and the use of

simulated work task situations; descriptions of a situation in which needs for information are triggered on users. This gathered data allows the experimenter to analyse not only final results such as number of relevant objects retrieved but also the processes that led to judgements of relevance. The analysis of the performance data gathered is usually done through the examination of the relevant metrics such as Precision and/or Recall [5], however analysing cognitive data such as the thought processes that led to the user-valuations of the documentation retrieved – *relevance processes* as we call them – may not be as straightforward.

In this article we propose a custom plotting technique which provides a novel approach to analysing both the relevance and interaction information gathered using Borlund’s method. This approach involves customised visualisation techniques as well as the usage of protocol analysis. Qualitative data such as verbal reports are transformed into quantitative data using protocol analysis techniques which include transcriptions, segmentation and tagging of the segmented transcriptions. Once tagged, the segments can be analysed using standard quantitative measures. A quick overview of the potentially emerging patterns is obtained using a custom plotting of the data. This plotting includes information about the dimensions of relevance, the sequence of relevance judgement processes and the interactions observed during the search sessions.

The remainder of the article is structured as follows. In Section 2 we introduce Barry and Schamber’s relevance criteria classes [2]. Section 3 describes think-aloud protocols and their processing. The main contribution of this work, namely relevance criteria profiles and session visualisations, are introduced and discussed in Section 4. In Section 6 we explore the data obtained from a user study conducted during the first half of 2008 using the techniques described in the previous section. We conclude with some final remarks and recommendations for future work in Section 7.

2 Relevance Criteria

Relevance judgements are often reduced to being binary judgements, or graded assessment, of relevance at best (cf. discussions in [3]) providing no explanation to why the value was assigned. It could be that while a user considers one document to be relevant based on the length and depth of the information provided, s/he considers another document relevant based on it providing factual data and it being well written. In this paper, we focus on some of the reasons that might motivate relevance judgements.

Robertson and Hancock-Beaulieu [10] refer to these cognitive and behavioural aspects and describe three revolutions: the cognitive revolution, the relevance revolution and the interaction revolution. Briefly, the cognitive revolution posits the need of realism in investigating the formation of information needs. The relevance revolution also requires realism but in assessing relevance. The interactive revolution is about interactivity and IR not being a single-query process but more of a query-read-refine one. These three revolutions were acknowledged by

Borlund in the method of evaluation for Interactive IR (IIR) systems presented in [4].

Relevance criteria are preferences expressed by users when evaluating whether to obtain and use information, i.e. when they are evaluating the relevance of said information. Barry and Schamber suggest that there is “*evidence that a finite range of [relevance] criteria exists and that these criteria are applied consistently across types of information users, problem situations, and source environments*” [2]. The starting point they suggest for examining relevance criteria consists the overlap of taxonomies resulting from two studies [1,11] on user relevance criteria. Both studies are similar in the methodologies used however the types of users, information sources and formats are quite different. In our work, we extend this overlap with some of the criteria appearing in Barry’s original taxonomy [1]. The extension includes three forms of information novelty, user’s background knowledge and their ability to understand the information. Some of the relevance criteria codes used are listed below:

- Depth/Scope/Specificity: whether the information is in depth or focused, has enough detail or is specific to the user’s needs. Also whether it provides a summary or overview or a sufficient variety or volume.
- Tangibility: whether the information relates to tangible issues, hard data/facts are included or information provided was proven.
- Affectiveness: whether the user shows an affective or emotional response when presented the information.
- Ability to Understand: user’s judgement that he/she will be able to understand information presented
- Document novelty: the extent to which the document itself is novel to the user

Here, we focus on profiling users and sessions with respect to their use of such relevance criteria in judging document relevance within the context of a task. We studied 21 subjects. These subjects were characterised by three types of affiliation (10 subjects from computing, 8 from information management, and 3 from pharmacy). Subjects were also grouped according to their levels of research experience (10 Ph.D. students, 7 researchers, and 4 senior researchers) and were assigned a task according to this level: writing a literature review for a thesis, framing the impact of a grant proposal, and preparing a keynote speech at a conference respectively.

By understanding relevance criteria usage (e.g. the frequency or distribution of selected criteria), and eventually understanding their relation to user interaction and their effect on relevance judgment, we might be able to determine which criteria to make explicit for what types of users within end-user services, and move towards a more comprehensive evaluation of retrieval system performance that takes the user’s cognitive process, interaction and tasks into consideration.

3 Talk-Aloud Protocols

Talk aloud protocols are based on the idea that talking aloud while solving a task provides a view of the thoughts as the task solving process is ongoing [6]. In an IR context using talk aloud protocols would provide a researcher with a raw view of the relevance judgement processes that users go through when searching for literature. By observing these processes, a researcher can examine them and in turn observe the relevance criteria within those processes.

After the verbal reports have been collected, they are transcribed and have to be segmented in utterance which are then to be encoded. The granularity of encoding performed on the utterances, if any, will depend on the researchers' needs. In our work we initially encoded utterances using one or more labels from the following encoding:

- Interaction: any utterance that indicates the participant is performing an operation on/with the system or interacting with it, e.g. reading a document, clicking on a document surrogate, going back a page, etc.
- Intent: any mention of the participant's intentions regarding the obtained information or regarding their actions, e.g. using a retrieved document to impress their supervisor or initiating a search in the hopes of finding a particular type of information.
- Relevance Criteria: any mention of factors that may affect the participant's choices regarding whether they are to keep or not a document, e.g. if the user picks the document because it is a survey.

Utterances encoded as *interaction* were further encoded according to the following listing:

- Navigation: user interacts with the system by navigating, e.g. closing a document window, going back a page, etc.
- Reads out loud: user interacts with the system by reading a portion of text out loud

and utterances tagged as *relevance criteria* were encoded using the taxonomy of relevance criteria described in Section 2.

4 Relevance Criteria Profiles

Relevance criteria profiles are constructed by aggregating and counting occurrences of relevance criteria as observed during a search session. As such they provide a global view of the occurrence of relevance criteria during the session. The visualisation technique rests on the “relevance criteria piles” metaphor. These piles represent relevance judgement processes. A relevance judgement process is then defined as the sequential use of relevance criteria as delimited by interactions. Visualising data using our method can help uncover potentially emerging patterns in the users's interaction behaviours, relevance criteria usage and even

potentially anomalous search sessions. Other studies related to relevance criteria have mostly concentrated on qualitative investigations (e.g. [11]) or simple statistics presented in tables (e.g. [12]). Our method, in contrast, aims to provide a more comprehensive view of criteria usage that will highlight patterns with respect to users and sessions.

Coded utterances are grouped at the session level and counted; all mentions of a particular relevance criterion within the search session contribute to a single count for that criterion. For any one participant there is what we define a “relevance criteria profile”. A relevance criteria profile is the grouping of the mentions of the relevance criteria during the search session. A typical relevance criteria profile, visualised as a chart, looks like Figure 2. These profiles provide a global view of the number of times that each criterion has occurred during the search session for each participant. To make the numbers comparable across profiles, we normalise the counts within each profile by dividing by the sum of all criteria mentions: i.e.

$$rc'_i = \frac{rc_i}{\sum_{j=0}^N rc_j} \quad (1)$$

where rc'_i is the new, normalised, count for relevance criterion i , rc_i is the count for relevance criterion i and N is the total number of relevance criteria (in this article $N = 15$).

Aggregating profiles, for instance by participant’s affiliation or research experience does not require any special processing. Criterion counts are added by restricting the sums and counts to the group for which the profile is being created.

Profiles can be further compared by using the Jensen-Shannon (JS) divergence measure [9] for comparing profiles as it is based on the Kullback-Liebler [8] divergence but is symmetric. The JS divergence considers the KL divergence between p and q under the assumption that if they are similar to each other they should both be “close” to their average. As the JS divergence is based on the KL divergence, the smaller the divergence the more similar the two profiles are. Normalised relevance criteria profiles satisfy the properties of discrete probability functions so they can be compared using this divergence measure.

5 Session Visualisation

As a complement to global relevance profiles we designed a technique for visualising search sessions. Graphs resulting from applying our technique include information on the order of occurrence of the relevance criteria observed during a search session and the recorded interactions (if there were any).

Sequence is denoted by a time line. The time line only denotes an order in time and not any measure of it; equal spacing on the line does not mean equal time spans in the session. Relevance criteria ordering and grouping are represented as piles of coloured blocks. Each block represents the observation of a particular relevance criterion. Different criteria are assigned different colours.



Fig. 1. An example with three relevance criteria and interactions plotted

With relevance criteria piles we model relevance judgement processes. As long as relevance criteria are observed together one after the other with no other utterances of a different type in between, e.g. interactions, we consider them to be part of the same relevance judgement process. Interactions are plotted in between relevance criteria piles.

To plot a search session first we group the tagged utterances in relevance criteria groups. For each group, we plot the first relevance criterion in the sequence at the bottom of the pile, the second on top of it one unit to the right and so on. Blocks are made as long as need be so that the final shape of the pile resembles a staircase. An example graph can be seen in Figure 1. In this graph there are two interactions to the left and one to the right of the relevance pile which are plotted as N to denote a *navigation* interaction.

There are assumptions behind the piles metaphor. First of all there is the assumption of aggregation. When a relevance criterion has been observed we assume that this criterion will apply all the way until the user has made a final judgement. The application of criteria is done sequentially until the user is able to make a judgement about the relevance of the information. The length of each block in the graph symbolises this assumption. One of the consequences, should this assumption hold true, is that the sequence in which criteria are used matters and that there might be a degree of relationship between relevance criteria. Users might follow a pattern when using relevance criteria. By using piles we can start analysing whether a user's relevance judgement process exhibits these dependencies between relevance criteria. We also assume that each criterion contributes, either negatively or positively, to a final judgement. Negative contribution are represented as a minus sign next to the block in the graph.

A second assumption is that we can isolate or delimit relevance judgement processes by the appearance of interactions. We observed that relevance judgements usually end with the user navigating away from the document. This interaction can be preceded by the explicit verbalisation of the relevance judgement, e.g. the user utters "I don't like this document". A pile is then defined as occurrences of utterances that are not interactions. There are, however, some shortcomings attached to these assumptions. First of all, depending on what the researcher considers to be an interaction, piles will (or will not) correspond to documents and their judgement processes as interactions are not necessarily all navigation interactions. Further encoding of interactions might alleviate this to a certain extent since the dynamics of the session might become more visible. Gathering click-through data and using it to better delimit the relevance judgement processes might also alleviate this situation.

Plotting sessions using our technique allows a researcher to investigate the relative strength, or importance, of a relevance criterion. In Figure 1 we see that

one of the three criteria mentioned has a negative sign next to it. This represents situations in which the user expressed a relevance criterion in a negative way, e.g. “this is too old, it’s from back in the 60’s”. In the example the criterion has been mentioned in a negative fashion, yet the judgement process continues. This may suggest that its strength, relative to the overall judgement process, is not as strong as to end it right there and then. The explanations can be varied, however the point is that researchers can direct their attention to further investigate these scenarios.

Choosing a Colour Sequence. According to Ware[13] the effectiveness of coding using colours for coding is degraded as more categories are added. Ware recommends 12 colours which are normally used when labelling using colours. The first six colours, which also correspond to the basic colours in the colour opponent theory[7], are: white, black, red, green, yellow and blue. The remaining six colours are: pink, grey, brown, magenta, orange and purple.

Taking the colours as an ordered sequence of recommendations, we use the number of occurrences of relevance criteria, in an aggregated profile, as indices to select an appropriate colour. The most occurring relevance criteria is then assigned the first colour in the sequence, the second most occurring criterion the second colour in the sequence and so on. The rationale behind this procedure is that, since aggregated profiles are obtained by averaging across users, higher relevance criteria counts mean that users have mentioned the criterion, on average, more often hence it is likelier to be observed in any one search session. Choosing the most contrasting colours for the most commonly occurring relevance criteria should make easier the visual detection of the different criteria.

6 Results

In this section we present and discuss data obtained from a user study carried out from January to August of 2008. A total number of 21 people accepted the invitation to participate in the study. All users were research scientists and were affiliated to one of three groups: the School of Computing, the Information Management Group and the School of Pharmacy. The main characteristic of the search task given to users was that it required them to search outside their research field for literature related to their own area of research.

6.1 Comparing Relevance Profiles

The global profile, aggregated from all the individual profiles, is depicted in Figure 2. We can immediately observe that *tangibility* and *depth/scope/specificity* are the most mentioned criteria. Relevance criteria profiles can be plotted together however before doing so they have to be normalised as described in Section 4. In Figure 3 the profiles of the three schools are plotted together. By plotting the profiles together we can quickly see similarities and differences. In the figure we see that while participants from the School of Computing have a distinguishable preference for tangible data, members of the other two schools prefer other

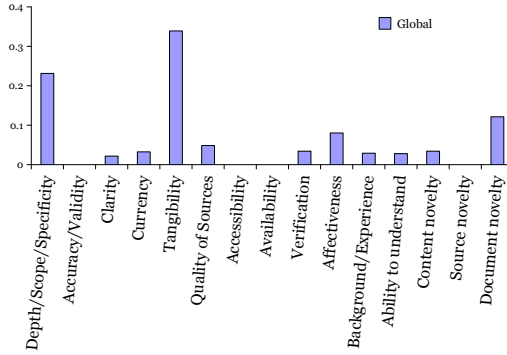


Fig. 2. Global aggregated relevance profile

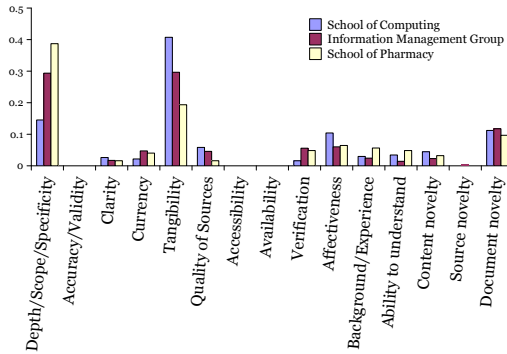


Fig. 3. The school profiles plotted together

aspects of the information such as its depth, scope and specificity. Furthermore, we can also observe that members from all three schools share the same interest (in terms of proportions) for the novelty of the documents found.

By plotting the divergence scores between all participants’s profiles and each other we can spot outliers but also see if there are any naturally emerging groups. The JS divergences between each individual profile and the other profiles are depicted as a matrix in Figure 4.

In each matrix, the value in cell (i, j) corresponds to the JS-divergence value between the profiles of participants i and j . Rows and columns are ordered by date in which the participant took part of the study. This leads to the participants being ordered by school, i.e. index values from 1 to 10 represent the School of Computing, from 11 to 18 the Information Management Group and from 19 to 21 the School of Pharmacy. The matrices in each map are all equal and the only difference between maps is the number of colours used as palette for the JS-divergence values; the redder the colour of the cells the less divergent the two profiles are. In all matrices, the profile in row/column 6 has a high divergence

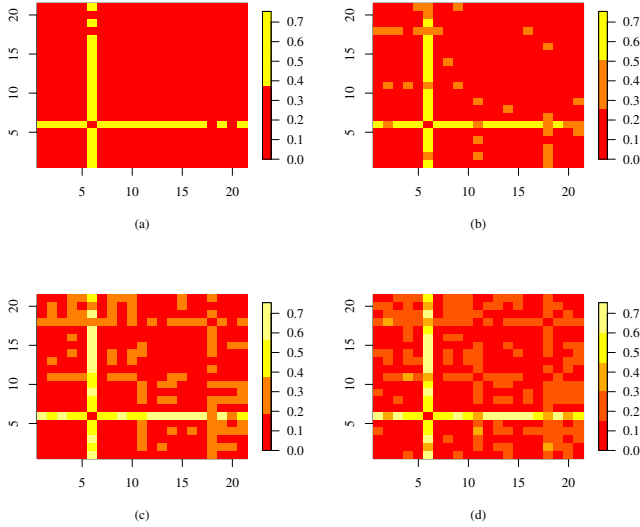


Fig. 4. Jensen-Shannon divergence measure between all individual profiles and the global profile

with almost all the other profiles. This suggests that the participant represented by the profile in row 6 is an outlier. In the last heat map, Figure 4 (d), we can observe that the profile in row 18 diverges with practically every other profile but with two. One of these two profiles is that in row 11 which also seems to diverge with most other profiles. In the figure we can also observe that the profiles of the participants of the School of Computing remain fairly convergent and that they diverge more with the profiles of the members of the School of Pharmacy than with those of the Information Management Group. The profile in row 17 seems to be very similar to almost every other profile with the exception of two: profiles in rows 18 and 4. There seems to be a group of profiles that are convergent, to a certain extent, with almost every other profile. These profiles are those in rows 1,2,3,7 (members of the School of Computing) and 12 and 17 (members of the Information Management Group). That these profiles are convergent with most other profiles could be due to that the participants represented by these profiles follow a globally shared behaviour in using relevance criteria to judge the relevance of the information presented, however before confirming/rejecting this suggestion, a closer inspection to the search sessions should be conducted.

6.2 Plotting Sessions in Practice

A much quicker approach to confirming the anomalous behaviour of the diverging profile found in Figure 4 would have been to look at the visual representation of the participant's search session. This visualisation is presented in Figure 5. At first sight it can be seen that the participant not only did not mention relevance

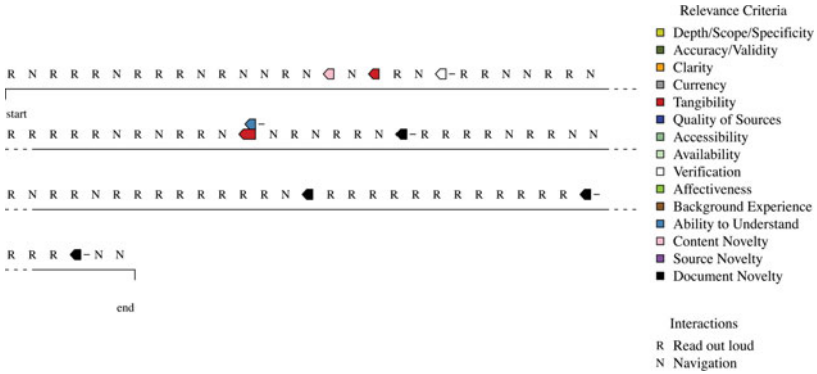


Fig. 5. The anatomy of an anomalous search session

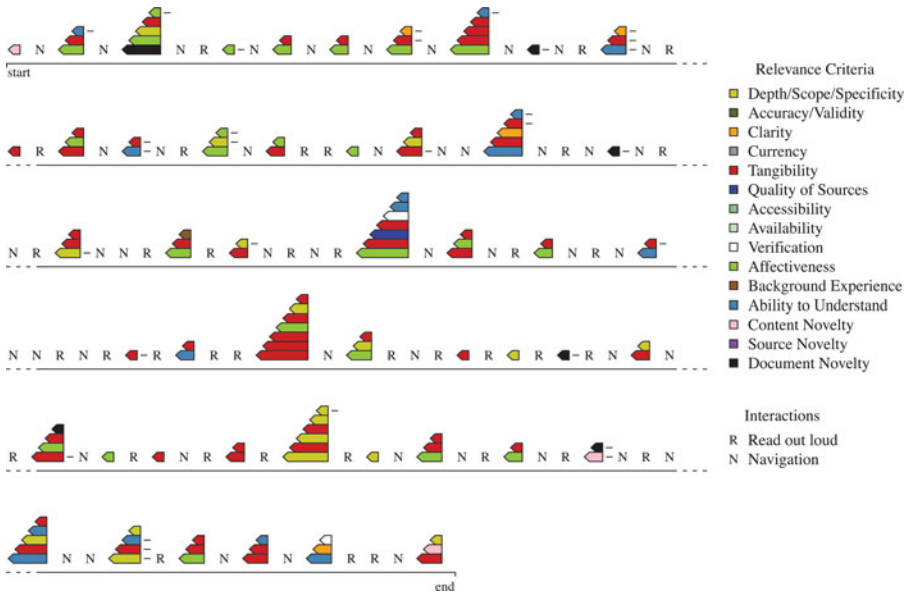


Fig. 6. A typical search session visualised using the piles metaphor

criteria very often but also that the participant spent almost all of the session reading out loud. This could reflect a misunderstanding in the instructions for the study or simply that the participant did not find any documents that were even remotely interesting.

Participant 2 (Figure 6) is a research student from the School of Computing. At a glance, if we interpret the number of expressions of *affectiveness* as a measure of engagement, we can observe that the participant is engaged from

the beginning, and remains so throughout the session. These affective responses, are represented as blocks coloured in light green. Effectively, out of 49 relevance judgement processes (depicted as coloured piles in the graph) 22 (about 45%) contain at least one expression of *affectiveness*. Affective responses seem to be, however, more frequent at the beginning than closer to the end of the session. Additionally, *tangibility*, which includes topicality, seems to play an important role during the participant's search session. Out of the 49 relevance judgement processes, 37 (about 75%) include at least one utterance encoded as *tangibility*. This complements the global view presented by the relevance criteria profile (see Figure 3) which showed that *tangibility* was a commonly used criterion by participants from the School of Computing. During the participant's session, *tangibility* not only was a commonly used criterion, but also one that was present in most relevance judgement processes. Moreover, the criterion is present in relevance judgement processes of different complexities covering almost the full range.

7 Discussion

In this article we presented the notion of relevance criteria profiles and a novel technique to plot the interactions and relevance criteria mentions observed during search sessions. We demonstrated, by example, how these tools aid the analysis of data. First, we showed how aggregated relevance criteria profiles provide global views of different user groups' preferences. We also showed how plotting relevance criteria profiles together can help uncover both (dis)similarities in relevance criteria usage at a global level. Outlier detection as well as cluster analysis are two of the types of analysis that can be performed when JS divergence scores between pairs of profiles are plotted together. Second, the visualisation technique presented in Section 5 was shown to aid with the analysis of search sessions. Using the data gathered from participant 2 we described some aspects of the search sessions that can be observed. We suggested that the participant, as well as being emotive, pays special attention to tangible data.

Relevance criteria are not theoretical concepts, but rather tangible and operationalising them can potentially impact positively on search services. Operational estimations of the most observed criteria may be embedded in systems in an attempt to increase their performance in returning relevant information. If, and only if, we can measure them. *Tangibility*, may be approximated, for instance, by looking at the number of tables in a document, and *depth/scope/specificity*, by looking at the number of pages in a document (document length has been mentioned frequently as a relevance criteria). Relevance processes, and the intertwined interactions, may be used to model user search behaviours in an attempt to personalise and adapt the system to better accommodate the current information needs of users.

References

1. Barry, C.L.: User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science* 45(3), 149–159 (1994)
2. Barry, C.L., Schamber, L.: Users'criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management* 34(2-3), 219–236 (1998)
3. Borlund, P.: The concept of relevance in IR. *Journal of the American Society for Information Science and Technology* 54(10), 913–925 (2003)
4. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research* 8(3), 8–13 (2003)
5. Cleverdon, C.W., Mills, J., Keen, E.M.: *Factors Determining the Performance of Indexing Systems*. Design, vol. 1., Test Results, vol. 2. Aslib Cranfield Research Project, Cranfield, Lagland (1966)
6. Ericsson, K.A., Simon, H.A.: *Protocol analysis: verbal reports as data*. MIT Press, Cambridge (1993)
7. Hurvich, L.M., Jameson, D.: An opponent-process theory of color vision. *Psychological Review* 64, 384–404 (1957)
8. Kullback, S., Leibler, R.A.: On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86 (1951)
9. Lin, J.: Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory* 37, 145–151 (1991)
10. Robertson, S.E., Hancock-Beaulieu, M.M.: On the evaluation of ir systems. *Information Processing Management* 28(4), 457–466 (1992)
11. Schamber, L.: Users'criteria for evaluation in a multimedia environment. *Proceedings of the 54 Annual Meeting of the American Society for Information Science* 28, 126–133 (1991)
12. Wang, P., White, M.D.: A cognitive model of document use during a research project. study ii. decisions at the reading and citing stages. *Journal of the American Society of Information Sciences* 50(2), 98–114 (1999)
13. Ware, C.: Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications* 8(5), 41–49 (1988)

Application of Session Analysis to Search Interface Design

Cathal Hoare and Humphrey Sorensen

Computer Science Department,
University College Cork,
Ireland
{hoare,sorensen}@cs.ucc.ie

Abstract. Evaluations of search features used in digital library environments are generally results centric, focussing on the outcome of an evaluation - for example, the number of relevant documents retrieved - rather than garnering an understanding of why that result was achieved. This paper explores how search feature development benefits from user-centered evaluation. By examining the application of an established web analytics technique, *session analysis*, to the development of search features and interfaces, it will be shown that designers can better understand how users conduct evaluation tasks. The feedback provided by this technique allows for clearer evaluation of an interface and admits iteratively evolving designs that are based on empirical data.

1 Introduction and Background

Many descriptions of information seeking processes have been written - Fisher et al [1] collated seventy different descriptions alone. These models describe examinations of search environments, user contexts and tasks being undertaken to provide an understanding of how people find information. These models, like any other, provide a framework for explaining an observed phenomenon, provide a common vocabulary and boundaries for collaborative investigations and, finally, provide a prediction of how a change in the environment would affect the outcome described by the model. This final point is an invaluable tool for those developing features and tools that support information seeking; as well as measuring the outcome of a task, those undertaking an evaluation can use a model of the participant's use of the test feature to understand how the user arrived at an outcome, and, to determine whether or not the outcome was the result of the intended use of the interface feature being examined. However, these models are underutilized in evaluation work; when the authors conducted an unscientific examination of full papers submitted to ECDL in 2009 [2] it was noted that, of the nine describing novel interactive information seeking features, only two papers presented a description of the features' patterns of usage.

Understanding how a user employs an interface is a common practice in many Web 2.0 type development methodologies. Driven by the desire to make commercially correct decisions, sites such as Amazon and Google, have developed



Fig. 1. Example of a funnel report in Google Analytics

evaluation techniques to understand how users navigate their sites. While a user is free to use the site in any way they choose, they invariably follow a pre-determined path to complete their purchase. For instance, an e-commerce site would have a process that starts with a landing page, usually describing special offers (for the unfocussed browser who might be tempted by a serendipitous offer) and providing a top level faceted view of the product categories. The next step of the process would encompass many views of product pages where the user locates a set of candidate products. This step is followed by a user making their purchasing choice and a progression to the checkout. The process then moves to support the user in making their purchase - collecting their payment details, delivery details and informing them of the terms and conditions of their purchase. The process completes by presenting the user with a clear conclusion to the process and offering them a set of options for continuing their exploration of the site. By *designing-in* the process, users are provided with tools that encourage particular strategies and lead to a successful conclusion for both the visitor and the site; tools and features are never added unless it can be shown that they improve patrons' adherence to a process. Despite this, users often fall out of the process for a multitude of reasons; they may never have intended to purchase, may have been dissuaded by adverse reviews or may not have had trust in the vendor through unclear terms and conditions - e-commerce sites need to understand why a purchase wasn't made.

Kaushik^[3] describes a web analytics^[1] report called a funnel report, an example of which is shown in Fig. 1; this report, so called because of the inevitable cone shape that arises from whittling down the initial population of users to the smaller group that completes the process. Information about a users traversal of an e-commerce site serves two purposes; The chart shows both the number of patrons that continue to the next step of the process and those that abandon their purchase. The next destination of those that abandon the search is shown as it can offer insight into why the search was abandoned. The linear nature of the chart reflects the nature of the e-commerce site it represents. Once common

¹ Web analytics is the measurement, collection, analysis and reporting of internet data for purposes of understanding and optimizing web usage. Many products, exemplified by Google Analytics, provide reports to facilitate this measurement.

points of abandonment are identified, targeted usability studies on the causes of breakdowns in process can be undertaken and solutions applied. The second use of traversal information is to allow more accurate application of artificial intelligence techniques, admitting recommender systems such as those provided by Amazon's A9 product engine.

2 Creating an Evaluation Model

This paper will describe an equivalent of the Funnel Report that was developed to understand user search processes. The analysis technique provides a model of states and transitions that users traverse during a search session. When a new search feature is developed, it is possible to express its anticipated and real affect on user behavior as a set of traversals; designs can be validated by comparing these sets. Similarly, different approaches can be compared by examining their respective models. The paper will explain how models are produced, before describing an application of this technique to a search feature comparing the feedback from this technique to the use of traditional measures, such as precision and recall, alone. The paper will conclude by discussing two pieces of future work derived from the model.

Fisher et al [1] described a plethora of user models. Broadly, these models can be partitioned into two sets. Some models describe specific tasks and behaviors and from these, derive a set of desired features that an interface should support [4]. Other models describe the interactions between all elements of the model as a process, showing transitions between states in the model; reference is made to beneficial and detrimental characteristics of these states and transitions [5] [6]. It is this type of model that can be used to evaluate the effect of a novel interface feature. When a transition map is produced by modeling a current interface and an interface that includes a novel feature using A/B testing, it can be determined if the new feature promotes anticipated transitions. If, when the resulting transition map is combined with metrics such as precision and recall, the feature is shown to have promoted improved retrieval metrics and produced the expected transition map, its design is validated. However, if the retrieval metrics are improved, and the transition map is not as expected, then the results were achieved for unexpected reasons; conversely, the expected transition map may arise while no improvement is seen in the retrieval metrics - in either case, despite potentially positive retrieval metrics, the design needs to be reassessed.

Several models were available to choose from; the authors chose a generic and expressive model to capture the wide range of scenarios that might have to be described; this provision better admits comparisons between evaluations. A general seeking model, described by Marchionini [6], details processes conducted at a query cycle level. In order to show the context and responsibilities of the stake holders in search, the general model is re-arranged as shown in Fig. 2. The user was made directly responsible for three states: recognizing the information need, defining the problem and reflecting on the state of search. The system was made solely responsible for query execution. The user is informed by the

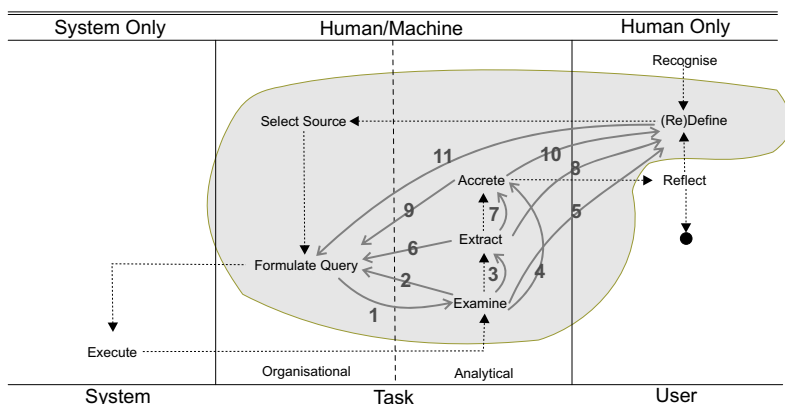


Fig. 2. Re-arranged General Model

system in four processes: source selection, query definition, result examination and information extraction. The authors add an extra process, *accrete*, to include any recording action, such as note-taking, annotation or bookmarking of salient information or insights learned during the search.

The model can be adjusted to take into account limitations of an experiment's setup. For example, in the experiment described in Section 3, source selection and query creation were grouped as a single task called formulate query, as there was just one source and only the transitions in the model that could be observed - those in the user/system category - were considered.

The model is initially represented as a directional graph where nodes represent states in the search process and edges, representing transitions between states, are weighted at zero; each time a transition is traversed, it's edge representation is strengthened by a single unit. Transitions are modeled as follows: each time a new query cycle is initialized, the transition between formulate query and examine (labelled 1 in Fig. 2) is strengthened. When a query is abandoned, the redefine to formulate transition (labelled 11) is strengthened. If a user invoked a search through the search field or a contextual search using information discovered on the result list itself, the transition between examine and formulate query (labelled 2), along with the return transition (labelled 1) are strengthened. If a document identified in the result set is viewed then the examine-extract transition (labelled 3) is strengthened. If a document or information identified in the result set is recorded, the examine-accrete transition (labelled 4) is strengthened. Finally, if the user abandons a query cycle while viewing the result set, then the examine-redefine transition (labelled 5) is strengthened.

Once a user views a document, they are in the extract state. If a user employs some information from the document as a new query or invokes a contextual operation from the document's result, the extract-formulate query transition (labelled 6) and transition 1 are strengthened. If viewing a document results in a document or information from a document being recorded, the extract-accrete

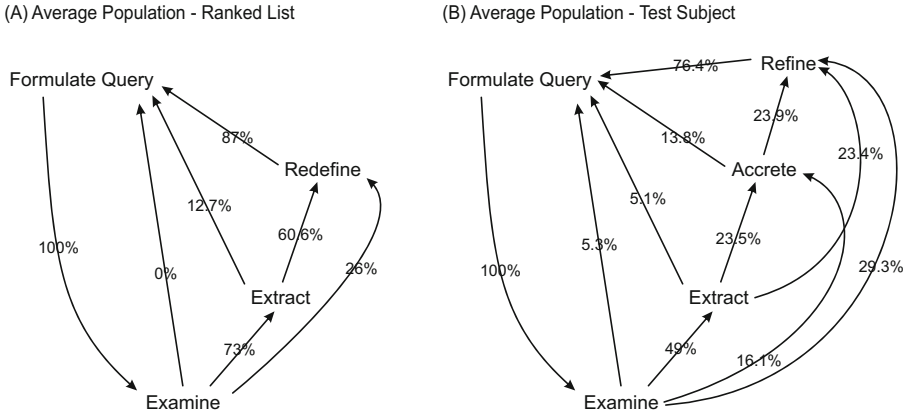


Fig. 3. Model Development without (left) and with (right) the Scrúdú feature

transition (labelled 7) is strengthened. Finally, if a query cycle is abandoned after viewing a document, the extract-redefine transition (labelled 8) is strengthened.

3 Applying the Evaluation Technique

Several information seeking features developed by the authors were modeled using this technique. The results of these evaluations are described elsewhere [7]. However, to better describe the benefits of this evaluation technique, one of the experiments will, briefly, be described and the implications of the results will be discussed. The experiment sought to determine the influence of adding a contextual document set to a ranked-list style search interface. The set allows users to add salient documents and notes found during an information seeking session. The set persists over the duration of a seeking session, and can encompass numerous search cycles. Further queries can be formed from the recorded artifacts through the use of contextual search operations. It was proposed that this arrangement would improve recall, a key requirement in exploratory search support systems [8] and better support the *orienteering strategy* [9] while providing a familiar interface design. If this proposition was true, the authors predicted an increase in recall figures and a reduction in the number of redefine-formulate traversals. Further details of the component, called *Scrúdú*, are provided in Appendix A.

A lab-based within-user instrumented study was conducted to evaluate the proposal. Two interfaces, a Google like control interface, and the test interface, identical to the first but with the addition of the *Scrúdú* feature, were evaluated using two tasks that were judged to be similar. By being consistent with collections and tasks, comparisons between features are admitted. The tasks and collection used were part of the TREC Complex Question Answering (CQA) track. Participants were chosen at random from a pool of email respondents.

After a period of familiarization with the interfaces, users had ten minutes to complete each task. Participants' interactions with each interface were recorded using the iShow video screen grab application. Interactions with the system were also logged to provide data for developing a tool that automates the analysis.

Analysis was conducted by reviewing each user's video and classifying each transition according to the model. These results were aggregated to produce the results shown in Fig. 3. Each transition is shown as a percentage of the total number of transitions made. The results are statistically significant ($t = 4.5$, $df = 14$, $.001 \leq P \leq .0005$) and for a 90% confidence level the error rate was ± 5.11 . The most common transition on both interfaces is redefine-formulate query. The Control scores 87% on this transition, while Scrúdú scores 76.4%. This implies that using Scrúdú, there is a greater tendency to formulate queries from information discovered in the examine (0% Control, 5.3% Scrúdú), extract (12.7% Control, 5.1% Scrúdú) and accrete (13.8% Scrúdú) states. These transitions can only be made if information from a result set informs the query. The level of recall achieved by the test interface was also higher. These two metrics allowed a causal effect to be drawn between the test feature and the expected result of its use.

The model reveals that, in this case, the expected usage pattern developed. However, other unexpected results were also observed; 0% of transitions made using the control were along the examine-formulate edge while the corresponding figure in the test interface was 5.3%. As this pattern was investigated it was determined that users began to use contextual search operations, such as similar documents, on the result list, having started to use similar features on the Scrúdú document set. This insight might prove prescient to a future evolution of the search feature.

4 Future Work and Conclusions

In order to allow the development of this feature, and to facilitate large group evaluations of features, an automated evaluation tool has been developed. By instrumenting an interface component, remote user actions can be captured and logged. The evaluation tool takes these logs, and by applying the rules of the model, described in Section 2, produces a traversal model of the users' actions. This tool can be used in realtime to trigger recommender system interventions or can be used to analyze the behavior of large numbers of users. The accuracy of the tool matches that of the analysis carried out by exhaustive manual examination of users' actions - each time an evaluation using the model is run, the users' actions are logged and the results of the tool's output are compared to the manual analysis. The manual analysis, of course, provides a finer grained insight to user actions and where possible it will continue to be effected.

This paper has highlighted the drawbacks of results oriented evaluation of information seeking tools. It has described a model of user states and transitions, and shown how maps of users' traversals, produced by this model, can provide a deep insight into how users arrived at the results. This insight is invaluable

for developers as it provides an understanding of how an interface is being used and allows accurate predictions of how changes to the interface might change users' behavior. The model also allows for deep comparisons between designs. The paper concluded with a description of how the model can be used to detect detrimental user strategies and proposed future work that would create system interventions to alleviate the impact of these.

References

1. Fisher, K., Erdelez, S., McKechnie, L.E.F.: Theories of Information Behavior. ASIST Monograph (2005)
2. Agosti, M., Borbinha, J.L., Kapidakis, S., Papatheodorou, C., Tsakonas, G.: The 13th European Conference on Research and Advanced Technology for Digital Libraries, Corfu, Greece (2009)
3. Kaushik, A.: Web Analytics: An Hour a Day. Wiley Publishing, Chichester (2007)
4. Bates, M.: The design of browsing and berrypicking techniques for the on-line search interface. *Online Review* 13(5), 407–431 (1989)
5. Blake, C., Pratt, W.: Collaborative Information Synthesis I: Models of Information Behavior of Scientists in Medicine and Public Health. *Journal of the American Society for Information Science* 57(13), 1740–1749 (2006)
6. Marchionini, G.: Information Seeking in Electric Environments. Cambridge University Press, Cambridge (1995)
7. Hoare, C., Sorensen, H.: The Enhanced Ranked List. In: Artificial Intelligence and Cognitive Science Conference, UCD, Dublin (2009)
8. Marchionini, G., White, R.W.: Find What You Need, Understand What You Find. *International Journal of Human-Computer Interaction* 23(3), 205–237 (2007)
9. White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. Morgan & Claypool (2009)

A Scrúdú in Detail

The Scrúdú² feature is shown in Fig. 4, where typical facts are enlarged. It is designed to support a wide range of search types, from simple lookup searches to complex exploratory search tasks of indefinite duration. The interface is made up of four distinct regions. The traditional query field, labelled 1 in Fig. 4, is the starting point for each search session. It supports full Boolean query syntax; this syntax is based on Google's format. On submitting a query, a result header, labelled 2, provides metadata for both the query and the result set. The result set returns a list of results - ten results per page. Each result shows a header (labelled 3), first line of the document, document metadata and a set of contextual operations relevant to the document type. The header can be clicked on to open the document in a new tab, or moused over to provide a popup of the document in-situ on the results page. The format of the result is intentionally formatted in the Google style to provide a sense of familiarity to users of the application. Some of the contextual operations act on the documents; while the *similar*

² Scrúdú is the Gaelic word for *examination, screening or inspection*.

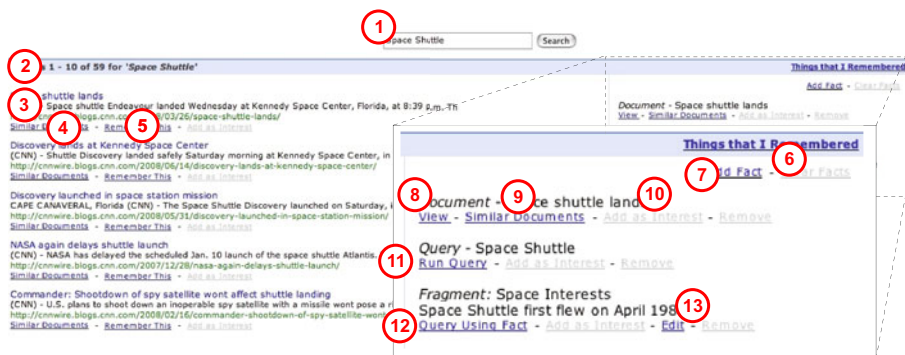


Fig. 4. Scrúd - the interface that was tested

documents (labelled 4) operation uses the document as query that searches the entire document collection. Other contextual actions provide interaction with the *fact list*.

The *remember this* action, labelled 5, allows a user to add a document to the *fact list* - titled *things that I remembered* on the interface. Items added to the list persist for the duration of the browsing session or until the user clears the list (labelled 6). To solve these issues, three types of facts are implemented by the interface. Besides documents, both queries and user-defined/metadata facts can be added as items in the fact list. Document facts can be added to the fact list in one of two ways. The first involves dragging the result onto the list; the user can also click on the *remember this* contextual action (labelled 5). As a result, on the fact list, a range of contextual actions (labelled 8) are available, including *view* (to view the document), *similar documents* (to invoke a search of the entire collection for similar documents, labelled 9) and *add to interests* (in order to persist the result beyond the current search session, labelled 10). Document facts can also be removed from the list. Query facts represent entire result sets. They are added as a fact by dragging the result set description (labelled 2) onto the fact list. This type of fact has three associated contextual actions including *run query* (labelled 11) which reruns the query. Query facts can also be added as an interest and can be removed from the fact list. User facts are used to capture information relevant to the user's current context, or to capture metadata about a particular document or class of documents. They are created by clicking *add fact* (labelled 7). This type of query can be used as a query (labelled 12). Seekers can use this type of fact to record significant authors, important dates and relevant links. User facts can be edited as new information becomes available (labelled 13). All facts can be annotated to note it's context and significance.

An Analysis of the Evolving Coverage of Computer Science Sub-fields in the DBLP Digital Library

Florian Reitz and Oliver Hoffmann

Dept. of Databases and Information Systems
University of Trier, Germany

reitzf@uni-trier.de, hoffmann@dbis.uni-trier.de

Abstract. Many scientists and research groups make use of the DBLP bibliographic project collection in various ways. Most of them are unaware of its internal structure, although it can have significant influence on their results. Prior work has shown that the collection does not cover all sub-fields of computer science in the same quality but has not provided an explanation for these differences. We introduce an extension of the DBLP data set which gives us a detailed picture on how DBLP has evolved since 1995. We show that the project started with a narrow focus on two sub-fields and discuss how additional themes have been added in recent years. We analyze the relations between sub-fields at different times and provide a model which explains the differences in coverage.

1 Introduction

The DBLP bibliographic project¹ is a frequently used collection of more than 1.3 million meta data records for publications in computer science and related fields. It indexes a significant part of the digital libraries of ACM, IEEE Computer Society and Springer as well as several smaller ones. The collection is freely accessible and has a high data quality. Therefore, it has been subject to a number of studies aiming at different questions, for example analyzing the structure of our research community [1,2] and predicting the future performance of researchers [3] or the popularity of research topics [4]. It has served as a test case for new approaches, e.g. search and retrieval functions, social network visualization and graph evolution models. Apart from scientific interest, DBLP has become an important tool for measuring the performance of single authors or institutions. Therefore, it has a significant influence on the awarding of research fund grants or the filling of vacant positions.

Despite its importance, little work has been done to analyze the collection itself and its development since it was established in the early 1990s. The size of the record collection (Fig. 1a) and the number of conferences and journals (Fig. 1b) can be found in different publications and websites. These figures show a massive growth during the last years but give no information on how this growth has evolved. Laender et al. showed in 2008, that DBLP did not cover all sub-fields of computer science to the same degree [5]. This bias is relevant for all applications we mentioned above. For example,

¹ dblp.uni-trier.de

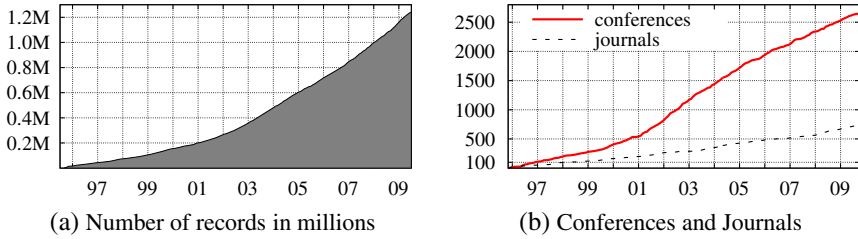


Fig. 1. Aspects of the evolution of DBLP since 1995

a scientist working in a poorly covered field will be underrepresented and important conferences or central papers might not be found in the collection.

To explain the differences in coverage we analyze how DBLP has evolved in the last 15 years. Consider the *International Conference on Very Large Data Bases (VLDB)* and the *IEEE CS Conference on Computer Vision and Pattern Recognition (CVPR)*. Both are central to their respective sub-fields and were established before the DBLP project started. However, VLDB has been listed in the collection from the very beginning while CVPR was not added before April 2003. We show that this is not a coincidence but an example for a long-term process. DBLP started with a narrow focus on database technology and logic programming which has widened in the course of the years to incorporate large parts of computer science today.

In this paper we show how the coverage of themes by DBLP changed over time and how this change took place. Our analysis is based on a set of backups of the DBLP collection which were created between October 1995 and September 2009. In Sect. 2 we describe the information we can extract from these files and discuss limitations of our approach. In Sect. 3 we compute some basic figures which describe the growth of DBLP and are necessary to understand more complex relations. In Sect. 4 we analyze how sub-fields were covered at different times. We utilize and extend the theme framework by Laender et al. to show how the coverage evolved. Section 5 concludes our work with an analysis of relations between sub-fields at different times. We show that this information is relevant for the growth of DBLP.

2 Data Source

The internal representation of DBLP is a directory tree. For each conference there is a folder which contains a small XML file for each associated record. For example, the directory `/conf/sigmod/` contains all records of the SIGMOD conference. File name and directory path define a unique identifier for each record which is almost never changed once it is assigned. There is no version control system like *CVS* or *Subversion* which could provide a list of changes to these files. Only the last time of modification, which includes creation, is stored in the modification date field managed by the operating system. We are mainly interested in the time the file was created, i.e. when the record was added. The modification date field contains this information as long as the record is not changed. Between 1999 and 2009, 775,650 modifications affected about

60% of all files. Each modification overwrites the modification date so that it does not provide the time of creation anymore. Therefore, we consider backups of the DBLP tree. These snapshots were created on each day between October 1995 and September 2009 where there was a change to the collection in a way that the file modification dates were preserved. We merge the directory tree of all versions into one and adjust the file-name so that many records are represented by multiple files. The oldest of these files has the time of creation as modification date. We call this tree the *historic collection*.

Our approach does not cover the early phase of the project because a different format was used to store the records before October 1995. However, this *missing past* only affects a small number of entries. All early records appear in the directory tree during 1995. In 2002 a broken script damaged the modification dates of all records. In the backups following this point, many records which were added shortly before this event are missing. We restored those by comparing the content of the damaged records with records from older backups.

3 Basic Figures

A first analysis of the historic collection provides several basic figures on the growth of DBLP. They are important for the understanding of more complex aspects of evolution we will discuss later. We have already seen how the number of listed publications in the directory tree increased from 25 in October 1995 to almost 1.3 million in September 2009. The actual number of publications at the start of this timeframe is higher because of the *missing past* problem. At the end of 1995 the directory tree contained about 14,000 files including those that were missing at first. In 2002 Ley [6] published a similar statistic based on the publication count between June 1996 and June 2002. These figures comply to our findings. The expansion was not linear. Figure 2a shows the number of new records by year. While less than 50,000 new entries were added before 2001 this number increased to more than 150,000 in 2007. Not all papers were added in the same year as they were published. The overlay in Fig. 2a shows the percentage of *old* publications which was always higher than 50%.

DBLP differentiates between conference papers, journal articles and six other types of publications. Conference papers and journal articles dominate the collection with an aggregated share of more than 95%. Figure 2b shows that the types were differently affected by the increasing growth. Before the year 2000, there were about the same number of articles and conference papers. After that, the number of conference papers started to grow faster. In fact, almost the entire amount of additional new papers in those years were conference papers. Figure 1b shows a similar increase in the number of conferences at the same time. In 2003, the number of articles started to grow as well and in 2009 there were even more new articles than conference papers. For many conferences and journals additional publications have been added regularly since they were listed in the collection. However, some were discontinued after some time. Figure 2c shows the number of journals and conferences with no added records by year. In 2008, for example, 1544 conferences and 261 journals did not receive additional records.

Figure 2d shows that the number of authors did not increase in the same way as the number of records did. Since 2003, the number of new authors has been stable around 90,000. As a result, the average number of papers per author rises steadily.

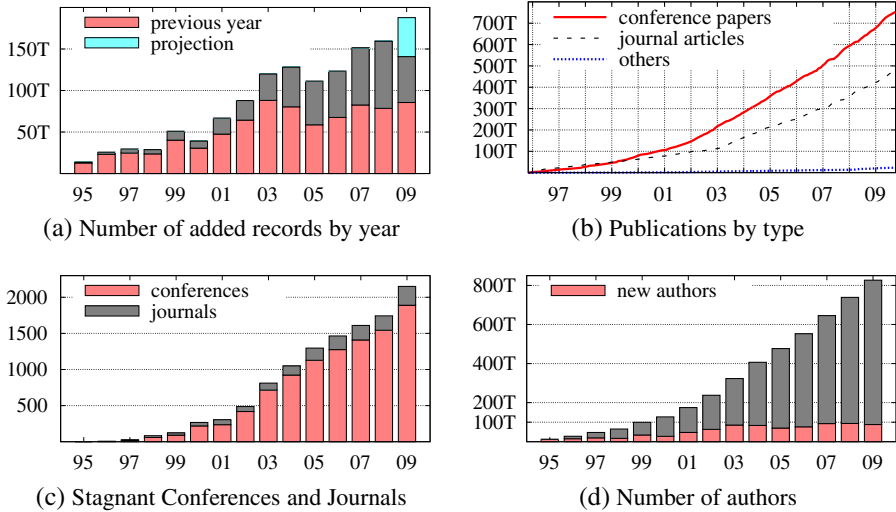


Fig. 2. Different indicators for the growth of DBLP between 1995 and 2009

4 Evolving Sub-field Coverage

We measure the coverage of a sub-field by how many of its related conferences are listed in DBLP. If a large number of these conferences are missing, we assume a low coverage. There is no common agreement on the definition of computer science sub-fields and which conferences are associated to them. There are a large number of lists available but most of them feature either a small number of fields with many conferences or a large number of fields with only two or three conferences each. In both cases we can not get significant results. In addition, it is often not clear how these lists were created so we can not rule out personal biases. In this paper, we use a thematic framework introduced by Laender et al. [5] in 2008 and refined by Martins et al. [7] in 2009. It features 27 themes (sub-fields) as listed in Table 1. Note that t_{17} , t_{20} and t_{24} are not used in the framework. In the version of 2009, 1000 conferences were assigned to the themes. For each theme t_i there is a set of associated conferences L_i . These sets were created by analyzing the publications of Brazilian computer scientists between 1954 and 2007. It was refined and completed by polls among researchers. As a result 27 themes were picked and the conferences were assigned to them. International conferences were preferred over local meetings and those with low reputation ratings were excluded. Journals were not considered. The size of L_i ranges from 12 (L_3) to 82 (L_{15}), i.e. no set is extremely large or almost empty.

In a first step, we determine which of the conferences were listed in DBLP in September 2009. We consider a conference as listed if at least one proceedings is covered. We obtain sets $D_i \subseteq L_i$ which contain the listed elements of L_i . With D_i and L_i we can define the coverage in September 2009 as $|D_i|/|L_i|$. The column *coverage* in Table 1

Table 1. The 27 theme groups based on Laender et al. [5] and Martins et al. [7] with associated figures

	Group description	L_i	D_i	coverage	missing date	duplicates	L_i^{1995}	L_i^{2000}	L_i^{2005}	G
t_1	Algorithms and Theory	42	33	78.6%	1 2.4%	1 2.4%	26	31	42	41
t_2	Databases, Information Retrieval, Digital Libraries and Data Mining	45	43	95.6%	1 2.2%	2 4.4%	24	35	45	43
t_3	Computational Biology	12	10	83.3%	0	0	3	7	11	12
t_4	Applied Computing	33	16	48.5%	2 6.1%	1 3.0%	15	25	33	32
t_5	Comp. Graphics, Image Processing and Computer Vision	69	52	75.4%	1 1.4%	0	36	54	66	69
t_6	Integrated Circuits Design	46	33	71.7%	0	0	34	45	46	46
t_7	Software Engineering and Formal Methods	73	60	82.2%	1 1.4%	2 2.4%	31	55	70	71
t_8	Geoinformatics	13	8	61.5%	1 7.7%	1 7.7%	6	10	12	12
t_9	Computer Education	24	9	37.5%	2 8.3%	0	12	18	24	24
t_{10}	Artificial Intelligence	49	39	79.6%	0	0	22	35	47	49
t_{11}	Human Computer Interaction	27	26	96.3%	0	1 3.7%	20	23	26	26
t_{12}	Programming Languages	41	36	87.8%	0	0	23	34	41	41
t_{13}	Multi-thematic	17	14	82.4%	0	2 11.8%	7	8	17	15
t_{14}	Operational Research and Combinatorics	28	9	32.1%	1 3.6%	2 7.1%	12	17	25	26
t_{15}	Comp. Networks, Distributed Systems and P2P Systems	82	67	81.7%	1 1.2%	3 3.7%	36	53	78	79
t_{16}	Simulation and Modeling	16	12	75.0%	0	0	10	11	16	16
t_{18}	Web and Multimedia and Hypermedia Systems	43	39	90.7%	1 2.3%	2 4.7%	12	27	42	41
t_{19}	Games and Virtual Reality	27	16	59.3%	0	0	4	13	24	27
t_{21}	Information Systems	14	11	78.6%	0	1 7.1%	7	11	14	13
t_{22}	Machine Learning	38	33	86.8%	1 2.6%	0	21	32	38	38
t_{23}	Robotics and Control and Automation	40	16	40.0%	1 2.5%	1 2.5%	24	28	37	39
t_{25}	Security	39	32	82.1%	0	0	15	24	38	39
t_{26}	Computer Architecture, High Performance Systems and Operating Systems	60	53	88.3%	2 3.3%	3 5.0%	38	45	60	57
t_{27}	Embedded, Real Time and Fault Tolerant Systems	25	22	88.0%	0	1 4.0%	11	18	24	24
t_{28}	Ubiquitous Computing	31	28	90.3%	3 9.7%	1 3.2%	2	10	31	30
t_{29}	Formalism, Logics and Computational Semantics	34	31	91.2%	0	0	23	29	34	34
t_{30}	Natural Language Processing	32	18	56.3%	2 6.3%	0	15	25	30	32

lists our results. They conform to the findings of Laender et al.. There are significant differences in the coverage of themes. For example, while almost all conferences from L_2 are contained in DBLP two thirds of L_9 are missing.

To compute the coverage for past times we need to know when a conference was established and when it was added to DBLP. The date of the first venue is important because the conference can not be listed in DBLP before that. We extract this information from digital libraries and conference websites. If meetings split, merge or change their name, it becomes difficult to tell when they started. When we could not find information on the first venue we assumed a start in 1995. Table II lists the number of missing dates per theme. For many conferences got only the year of the first venue. We define L_i^y as the subset of conferences from L_i which were established at the end of year y . We use the historic collection to find the date a conference was added. We have to consider that DBLP and Laender et al. use a different granularity to determine what an independent conference is. For example the *International Middleware Conference* is listed by Laender et al. as well as the associated workshops *MPAC* and *MGC*. However, all three entities share the DBLP key `conf/middleware`. With only one key, it is difficult to map list entries and papers and tell when they were added. The column *duplicates* in Table II shows the number of affected conferences per theme. Again, we choose the earliest reliable date. Similar to L_i^y , we obtain sets D_i^y for the listed conferences. With this information, we can define the coverage in year y as

$$cov : Theme \times Year \mapsto [0, 1] \quad cov(t_i, y) := \frac{|D_i^y|}{|L_i^y|} . \tag{1}$$

Note that *cov* does not monotonically increase over time but can decrease when uncovered conferences become relevant for L_i^y . We do not weight our result by the size of a conference or an importance rating because this information is either hard to come by or subjective. We compute *cov* for all years between 1995 and 2009. The results we obtained for 1995 and 2009 have limited significance. The 1995 data is strongly affected by the problem of the *missing past*. For 2009, new conferences are missing in the framework by Laender. We also have to mention that the framework had already been published at that time and might have influenced the growth of DBLP.

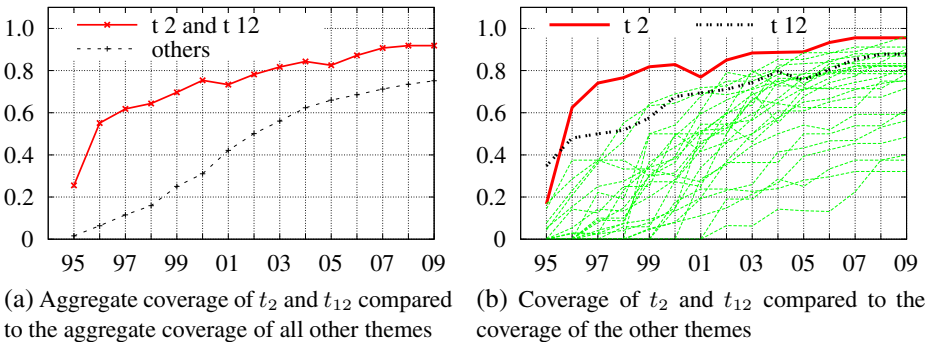


Fig. 3. The coverage of t_2 and t_{12} compared to the coverage of the other themes. (X-axis: year and the Y-axis: coverage)

As DBLP originally stood for *DataBases and Logic Programming*, we first consider the related themes t_2 and t_{12} which include logic programming as a sub-field. The coverage results (Fig. 3a) show that the aggregate coverage of t_2 and t_{12} has always been significantly better than the aggregate coverage of the other sub-fields. Figure 3b shows the individual coverage of t_2 and t_{12} compared to the individual coverage of all other themes. Except for 1995, the coverage of t_2 has always been the best in the collection. Only recently, t_{11} gained a similar coverage. t_{12} has not performed that well which may be caused by the fact that it incorporates not only logic programming and that many listed conferences were not in the original scope of DBLP. However, it has always been among the best covered themes.

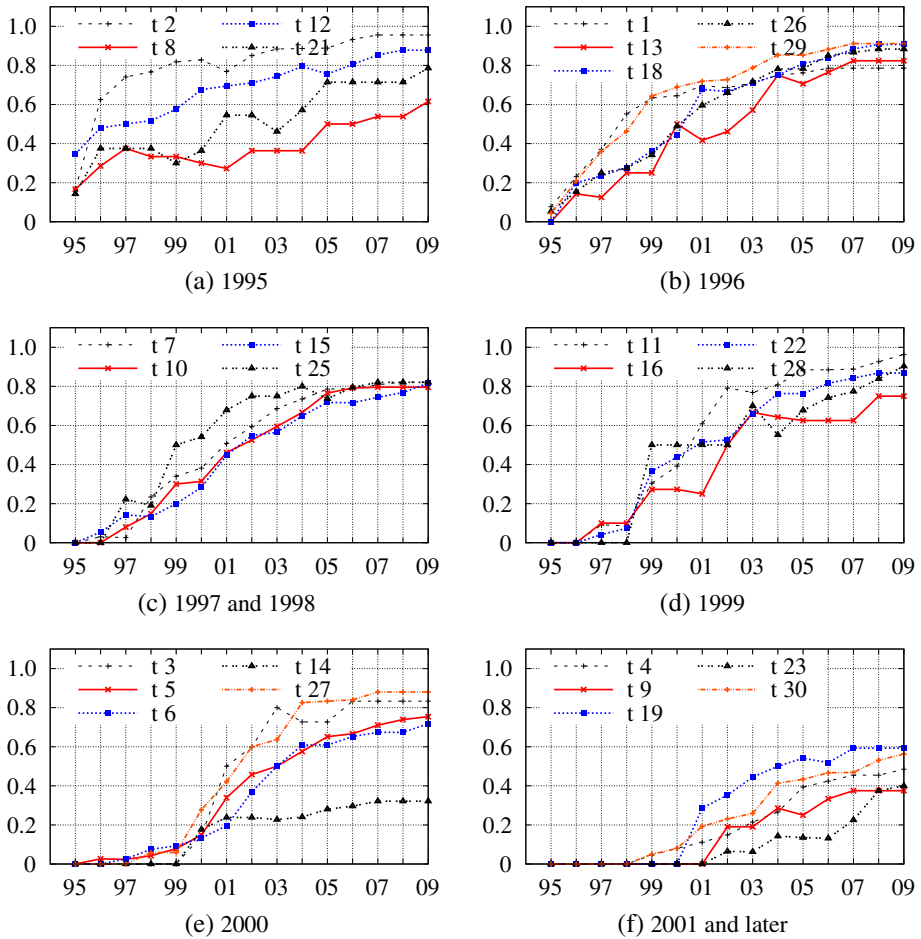


Fig. 4. Coverage values of different themes. The distribution is based on the first year the theme reached a coverage of at least 0.1. (X-axis: year and the Y-axis: coverage).

Figure 4 gives a detailed view on how the coverage of each theme has evolved. To improve the readability, we split the 27 themes into six groups depending on the year they first reached coverage of at least 0.1. Note that themes t_3 , t_8 , t_{13} , t_{16} and t_{21} have less than 20 conferences assigned to them. Small changes to L_i^y or D_i^y have a strong influence on the coverage. We must also take into account that the conference list concentrates on recent meetings. Those which were not continued after 2004 are rarely covered. Table 1 shows the size of L_i^{1995} , L_i^{2000} and L_i^{2005} . For several themes, L_i^{2000} misses a significant number of conferences. Nevertheless, these conferences are relevant for the coverage at some point in time and we discussed in Sect. 3 that many of them were added to DBLP. Because of this, the significance of the coverage values for long-past years is limited.

If we consider the coverage in 1998 we see that the scope of DBLP was still focused on t_2 and t_{12} . Aside from these themes, only t_1 had more than half of its conferences listed. Nine themes did not have any entry at all. The mean coverage at that point was 0.183 with a high standard deviation of 0.205 which also underlines the large differences. Contemporaneous with the increase of new records and conferences we discussed in Sect. 3, the coverage of several themes started to rise in the year 2000. The number of new papers for conferences already listed remained constant and most of the additional new records could be used to widen the scope. The increase of coverage for most of the themes in Fig. 4c and Fig. 4d was rapid compared to the improvement of themes listed in Fig. 4a and Fig. 4b. There is a direct relation between the year the threshold of 0.1 was reached and the final coverage value. All themes which started after the year 2000 have never reached a coverage of 0.6 or more. t_8 (Geoinformatics) and t_{14} (Operational Research and Combinatorics) have a lower coverage than other themes which started at the same years. Both themes lie on the edge of computer science which might give them a low priority for extension.

5 Relation between Sub-field Communities

The conferences are not isolated from each other. One relation between conferences a and b is their common community, i.e., the set of all authors who published on a and b . When we consider this relation for a newly added conference, we find that usually more than 30% of the authors of new records have been listed in DBLP before. We assume that this integration into the existing collection is one criterion for a conference to be added to DBLP. If a theme has a low coverage a new conference must integrate with conferences from other themes. We assume that an increasing coverage requires a good integration with other themes, at least at the beginning. To analyze this relation, we consider the common community of themes t_a and t_b which we define as the set of authors who published on conferences associated with t_a and t_b .

The common community relation defines a small weighted graph. It contains a node for each theme and an edge for each pair of themes which shared at least one author. We define the weight of the edge between t_a and t_b by the reciprocal of the Jaccard index $J(t_a, t_b) = \frac{|A \cap B|}{|A \cup B|}$ where A and B are the sets of authors of t_a and t_b respectively. Thus, themes with strongly overlapping communities are connected by a *short* edge while there is a greater distance between themes with a small set of common authors.

We use the historic collection to compute a sequence of graphs $G = (g_{1996}, \dots, g_{2009})$. Each single graph represents the common community relation at the last day of each year between 1996 and 2009. As described in Sect. 4, some conferences in Laender et al.'s list share the same DBLP key. In 11 cases, the conferences associated to one key belong to different themes. Because it is difficult to tell which publication belongs to which conference if they share the same key, we can not differentiate the involved communities. If we ignore this problem we get strong bounds between themes which do not exist in reality. Column G in Table 1 shows how many conferences are left for each theme after we removed these duplicate entries. In the graph g_y , we omit nodes for all themes which were not covered at all at the end of year y . The first graph which contains all nodes is g_{2002} . Beginning with g_{2004} the networks are complete but the edge weights are unequally distributed.

To analyze the structure of G and the role of each theme we compute how central it is for the network. The *betweenness centrality* [9] $C_B(v)$ is based on the assumption that only the shortest paths between two nodes are relevant. If a node v is on a large number of shortest paths between all pairs of nodes, it is central for the graph. More formal, we compute

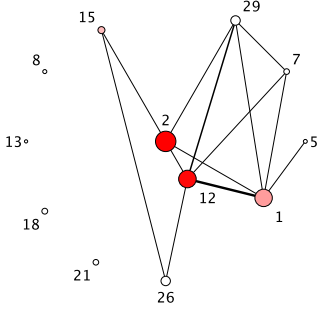
$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad \forall s, t \in V \quad (2)$$

where σ_{st} denotes the number of different shortest paths between nodes s and t and $\sigma_{st}(v)$ is the number of those paths which pass node v .

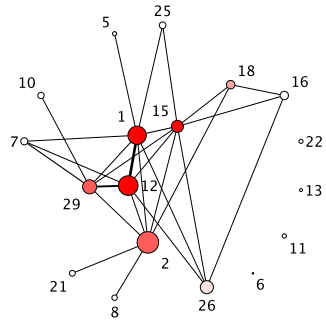
Figure 5 shows six selected graphs from G . To draw these graphs, we use a centrality layout based on the betweenness centrality which means that central nodes are positioned close to the center of the drawing. The lightness of the node coloring codes the C_B value. Dark colored nodes have a high centrality while white nodes have a betweenness centrality of zero. Nodes with a strong common community tend to be close to each other. The node area denotes the size of the respective community and the thickness of the edges the strength of relation. To improve the readability, we do not draw edges with a weight less than $thres_e$ of the maximum where $thres_e$ varies between 10% and 30%. All size information is relative to the respective graph.

We saw in Sect. 4 that there are a number of themes which were established early besides t_2 and t_{12} . However, except for t_1 and t_{15} none of them is central for g_{1996} in the sense of C_B . t_2 and t_{12} have similar centrality and community size but only a comparably small intersection. Both themes are connected to all other themes. While t_{12} has a smaller average distance to its neighbors ($d(\bar{t}_{12}) = 0.035$ and $d(\bar{t}_2) = 0.0033$), the standard deviation is much higher ($\sigma(t_{12}) = 0.048$ and $\sigma(t_2) = 0.0095$). While t_{12} achieves its centrality by a strong connection to a small number of other themes, t_2 has an even connection to the other nodes. These differences between the two central themes are valid until today.

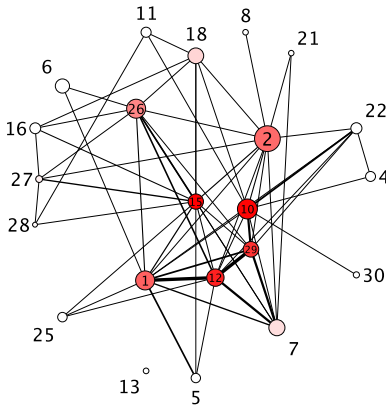
Prior to the burst of scope in the years between 2000 and 2002, the situation changed. In g_{1997} , we can see no most central theme but a cluster of four themes with a high betweenness centrality (t_{15} , t_1 , t_2 and t_{12}). If we consider the themes with a strongly increasing coverage in 1997 we see that all of them have a close relation to either t_{12} , t_{15} or both. This means that at that phase DBLP grew around these two themes while t_2 was less important. In g_{1999} this trend continued although t_2 played a more important



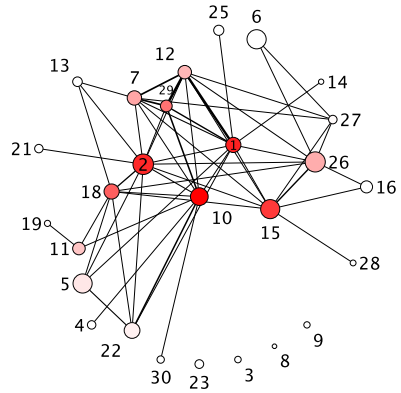
(a) 1996 $thres_e = 10\%$



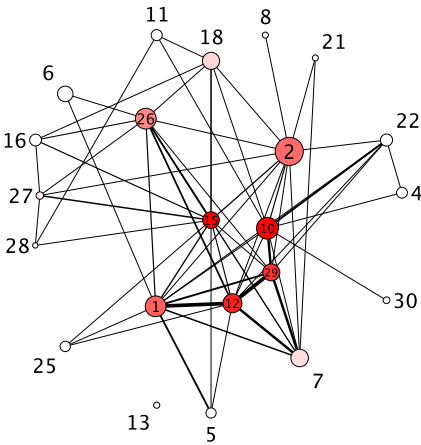
(b) 1997 $thres_e = 10\%$



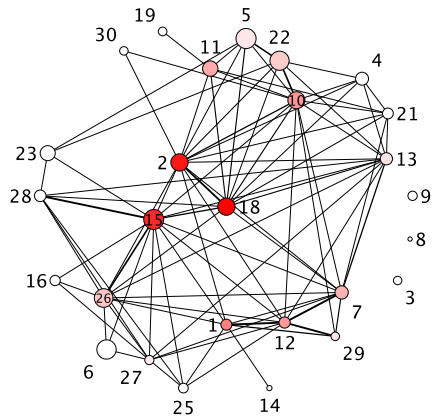
(c) 1999 $thres_e = 10\%$



(d) 2002 $thres_e = 20\%$



(e) 2005 $thres_e = 20\%$



(f) 2009 $thres_e = 30\%$

Fig. 5. Relations between sub-fields at the end of selected years

part for new themes then. We can see that there are no *thick* edges adjacent to t_2 unlike all other central themes. Again this is caused by the uniformity of the intersection sizes. t_2 does not connect strongly with one theme but it has the lowest standard deviation of edge weights of all nodes.

We can also see a heavily connected cluster consisting of t_1 , t_{12} , t_{29} and t_{10} which dominated the network in the following years. If we compare g_{1999} with the succeeding graphs g_{2002} , g_{2005} and g_{2009} we see that the basic structure is nearly stable though it lost significance. In 2009, there were 14 themes with a betweenness greater than zero which is a strong indication that the focus is on more than two themes now. We also see that t_2 has defended its central position while t_{12} has been marginalized.

When we consider themes with a weak integration in the single graphs we find that they are unlikely to have a significant increase in the near future. t_8 and t_{14} have a low coverage compared to other themes which started to become relevant at the same time and have always been poorly integrated. The respective sub-fields *Geoinformatics* and *Operational Research and Combinatorics* lie on the edge of computer science so it is difficult to integrate them with other themes. We assume that G obeys the *the rich get richer* rule which has been verified for a large number of dynamic networks. If we consider the lowest integration of all graphs we will find that this set is very stable. Between 2003 and 2008, t_3 , t_8 , t_9 , t_{11} and t_{30} were always the least integrated themes. They have in common that they are only loosely related to the original scope of interest. Assuming that an increasing coverage requires a strong integration we expect that these themes will have a low coverage in the future as well.

6 Related Work

As stated before, there is only little work on the structure and evolution of DBLP. The coverage of DBLP or the lack of it is mentioned in different papers but only a small number of studies deals with the question in more detail. We already discussed the work of Laender et al. who aimed at an performance analysis of Brazilian PhD programs. In 2005 Petricek et al. [8] compared DBLP with the citation database CiteSeer. Among other results, they presented two probabilistic models of how both projects acquire new records. They found that DBLP covers about 24% of all publications in computer science. However, they do not define the borders of *computer science* and give no details on sub-fields. Our results show a total coverage of 65% at the end of 2005 for the conferences listed by Laender et al. and Martins et al. which is only a small subset of all computer science conferences. However, by the way this list was created, we can assume that it contains the most relevant ones.

7 Conclusion and Future Work

Using the historic DBLP collection, we showed that there are thematic biases in the coverage of computer science by DBLP and how they evolved over time. We saw that the collection started with a small scope which gradually widened after the year 2000. Our analysis of the relation between the communities of the different themes showed

that a large number of common authors help increasing the coverage. Based on these findings, we made a vague prediction of the future development.

The most important drawback of our approach is that the theme lists contained a large number of conferences but no journals. Because of the way it was created it lacks information on short-lived meetings which might be relevant for the coverage at some time. Future work will have to find solutions to this problem. A thematic clustering might provide themes for a larger set of streams. However, prior to that, we have to solve the problem of finding sufficient data for streams not listed in DBLP.

Acknowledgements. We thank Alberto Laender for providing us with the conference list and Michael Ley for giving us feedback on the analysis.

References

1. Deng, H., King, I., Lyu, M.R.: Formal Models for Expert Finding on DBLP Bibliography Data. In: Proc. of the ICDM 2008, pp. 163–172. IEEE CS, New York (2008)
2. Elmacioglu, E., Dongwon, L.: On six degrees of separation in DBLP-DB and more. *SIGMOD Record* 34(2), 33–40 (2005)
3. Huang, Z., Yan, Y., Qui, Y., Qiao, S.: Exploring Emergent Semantic Communities from DBLP Bibliography Database. In: Proc. of the ASONAM, pp. 219–224. IEEE CS, New York (2009)
4. Li, X., Foo, C.S., Tew, K.L., Ng, S.-K.: Searching for Rising Stars in Bibliography Networks. In: Thou, X., Yokota, H., Denk, K., Liu, Q. (eds.) DASFAA 2009. LNCS, vol. 5463, pp. 288–292. Springer, Heidelberg (2009)
5. Laender, A.H.F., de Lucena, C.J.P., Maldonado, J.C., de Souza e Silva, E., Ziviani, N.: Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *SIGCSE Bulletin* 40(2), 135–145 (2008)
6. Ley, M.: The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In: Laender, A.H.F., Oliveira, A.L. (eds.) SPIRE 2002. LNCS, vol. 2476, pp. 1–10. Springer, Heidelberg (2002)
7. Martins, W.S., Gonçalves, M.A., Laender, A.H.F., Pappa, G.L.: Learning to assess the quality of scientific conferences: a case study in computer science. In: Proc. of the JCDL, pp. 193–202. ACM, New York (2008)
8. Petricek, V., Cox, I.J., Han, H., Councill, I.G., Giles, C.L.: A Comparison of On-Line Computer Science Citation Databases. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 438–449. Springer, Heidelberg (2005)
9. Freeman, L.C.: Centrality in social networks conceptual clarification. *Social Networks* 1(3), 215–239 (1978-1979)

Analysis of Computer Science Communities Based on DBLP


Maria Biryukov¹ and Cailing Dong^{1,2}

¹ Faculty of Science, Technology and Communications, MINE group,
University of Luxembourg, Luxembourg

`maria.biryukov@uni.lu`

² School of Computer Science and Technology, Shandong University,
Jinan, 250101, China

`cailing_dong@mail.sdu.edu.cn`

Abstract. It is popular nowadays to bring techniques from bibliometrics and scientometrics into the world of digital libraries to explore mechanisms which underlie community development. In this paper we use the DBLP data to investigate the author's scientific career, and analyze some of the computer science communities. We compare them in terms of productivity and population stability, and use these features to compare the sets of top-ranked conferences with their lower ranked counterparts. 

Keywords: bibliographic databases, author profiling, scientific communities, bibliometrics.

1 Introduction

Computer science is a broad and constantly growing field. It comprises various subareas each of which has its own specialization and characteristic features. While different in size and granularity, research areas and conferences can be thought of as scientific communities that bring together specialists sharing similar interests. What is specific about conferences is that in addition to scope, participating scientists and regularity, they are also characterized by level. In each area there is a certain number of commonly agreed upon top ranked venues, and many others – with the lower rank or unranked. In this work we aim at finding out how the communities represented by different research fields and conferences are evolving and communicating to each other. To answer this question we survey the development of the author career, compare various research areas to each other, and finally, try to identify features that would allow to distinguish between venues of different rank.

Research communities have been studied in a variety of perspectives. Since the late 90-th they are subject of social network analysis [\[9,7,11,4,6,2\]](#). Another

¹ This is a short version of the paper. The full version which provides a more detailed discussion of the topic, supported by the neat quantitative analysis, can be found at the Computing Research Repository (CoRR).

Table 1. Research Communities and Corresponding Top Conferences

Abbreviation Area	Top Conferences	Non Top Conferences
ARCH Hardware&Architecture	ASPLOS, DAC, FCCM, HFCA, ICCAD, ISCA, MICRO	
AI Algorithm&Theory	COLT, FOCS, ISSAC, LICS, SCG, SODA, STOC	APPROX, ICSS, SOFSEM, TLCA, DLT
CBIO Computational Biology	BIBE, CSB, ISMB, RECOMB, WABI	APBC, ICB, ISBRA, CBMS, DLS
CRYPTO Cryptography	ASIACRYPT, CHES, CRYPTO, EUROCRYPT, FSE, PKC, TCC	
DB Data Bases & Conceptual Modeling	DEXA, EDBT, ER, ICDF, PODS, SIGMOD, VLDB	IDEAS, ABDIS, ADC, WebDB, DOLAP
DMML Data Mining, Data Engineering, Machine Learning	CIKM, ECML, ICDE, ICDM, ICML, KDD, PAKDD	MLDM, IndCDM, ADMA, RES, IDEAL
DP Distributed&Parallel Computing	Euro-par, ICDCS, ICPP, IPDPS, PACT, PODC, PPOPP	
GV Graphics&Computer Vision	CGI, CVPR, ECCV, ICCV, SLD, SIGGRAPH	
NET Networks	ICNP, INFOCOM, ICN, MOBICOM, MOBIHOC, SIGCOMM	
NLIR Computational Linguistics, NLP, IR	ACL, EACL, ECHR, NAACL, SIGIR, SPIRE, TREC	
PL Programming Languages	AFLAS, CP, ICFP, ICPL, OOPSLA, PLDI, POPL	
SE Software Engineering	ASE, CAV, FM/FME, Soft FSE, ICSE, PEPM, TACAS	
SEC Security	CCS, CSFW, ESORICS, NDSS, S&P	SCN, ISC/ISW, ISPEC, ACISP, WISA
WWW World Wide Web	EC-web, ICWE, IEEE/WIC, ISWC, WISE, WWW	WEBIST, SAINT, WECWIS, ESWC, ICWE

point of interest is the topic development and distribution in scientific communities [11,12]. Yet another branch of investigation aims at quality evaluation of scientific venues [5,13,10].

Our work bears on the previous research in that it focuses on statistical investigation of the scientific communities. Its contribution consists in:

- extension of a framework for author analysis in order to build a comprehensive profile of the researchers on DBLP;
- setting up and analysis of criteria that allows for both between-area comparison and comparison of conferences that belong to different levels, in an attempt to build up a framework for automatic evaluation of scientific venues.

This paper is organized as follows: in Section 2 we elaborate on the data collection. Section 3 is devoted to the author profiling. Section 4 focuses on the comparison between various communities and venues. Section 5 concludes the paper.

2 Data Collection

We use computer science bibliographic database DBLP to conduct our investigation. The database is publicly available in XML format at <http://dblp.uni-trier.de/xml/>. We downloaded the file in August 2009 and used conference publications for corpus construction. While DBLP covers 50 years of publications the data before 1970 is rather irregular. This is the reason why we consider publications from 1970 on.

As we are interested in a comparative analysis of different scientific communities and venues, we prepare two data sets: one represents the top conferences [2,6,13,10] with at least 10 years time span² in 14 areas of computer science, and another one, that consists of non-top conferences in 6 areas of computer science.

The conferences of the TOP and NONTOP sets are given in Table 1. Note that there are some differences between the two sets in terms of topical partitioning

² We have had to relax the “min 10 years time span” requirement when dealing with conferences in Computational Biology and World Wide Web because these are young areas that have started off at the end of 90s.

and number of covered subareas. This is explained by the fact that the data about the lower ranked conferences is less consistent and agreeable, and we have preferred to construct smaller though more reliable sets.

In these three sets above we use publications with the complete bibliographic record (0.948%) to build undirected co-authorship graphs that we use for our experiments in combination with other bibliographic data such as number of records, venue, year.

3 General Researcher Profiling

The authors in co-author network are typically investigated from the point of view of their contribution to the research. Thus particular attention is paid to the members of program committees [13], “fathers” of the influential research directions [12], authors with high citation index [8] or yet those researchers who get often acknowledged [5]. Such an approach yields an interesting but narrow image of the researchers community. In this section we aim at providing a broader view on the authors in entire DBLP and the areas described above by looking at their typical career length, individual performance pattern and publication distribution with respect to the top and non-top venues. Since our NONTOP dataset covers only a small part of the lower ranked venues listed in DBLP, we do not compare the TOP and NONTOP datasets to each other in this setting. Rather we contrast the data in TOP dataset to the global author statistics in DBLP.

3.1 Author Career Length

DBLP contains to hundreds of thousands distinct authors. But how many of them pursue a long scientific career?

Figure 1 gives a full account on the authors career length distribution among the various research areas in the TOP set, CS dataset, and DBLP as a whole. The chart 1a on the left represents percentage of authors with ≤ 5 career length, while the chart 1b on the right covers periods from 6 to 20 years. It turns out that top-ranked venues are dominated by authors with ≤ 5 years experience, and only $\approx 2\%$ stay publishing at top ranked conferences for more than 10 years. This is consistent with the figures obtained on the whole DBLP set: $\approx 1.4\%$ of authors have a longer than 10 years career. We hypothesize that the main component of DBLP authors is represented by PhD students who, after having finished their studies, leave the active scientific career. With respect to the research subareas, AT and CRYPTO have the lowest percentage of researchers with a short career and the highest percentage of people whose career length ranges between 10 and 15 years. The explanation lays probably in that fact that these domains require substantial mathematical background and thus time to obtain it which makes them harder to get in for the short time scientists, and more difficult for switching for those who spent so much time on it.

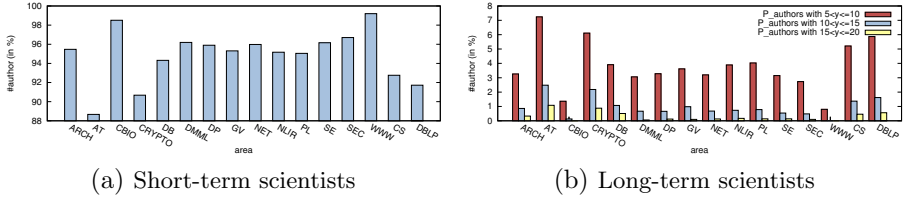


Fig. 1. Percentage of authors with $5 \leq \textit{career}$, and $6 \leq \textit{career} \leq 20$ years in Top set and entire DBLP

3.2 Individual Performance Pattern

We now turn our attention to the authors with ≥ 10 years experience since they are more probable to influence scientific community than “short time” researchers. In this subsection we investigate the author publication distribution over time and venues. For this purpose we distinguish between the following three groups of authors:

- Authors with ≥ 10 years experience of publishing in TOPset conferences and focusing on one area only;
- Authors with ≥ 10 years experience of publishing in TOPset conferences and focusing on multiple areas;
- Authors \in the TOPset with ≥ 10 years experience of publishing in the CS dataset, irrespective of the number of areas and conference rank.

The average number of publications produced by each category of authors per 5-years periods are plotted at Figure 2. The data reveals an interesting pattern: researchers in all three categories are much more active in the 2nd period of their career, and the single-area authors are even more active in the 3rd period. After that the productivity drops in the fourth period and remains stable with some minor fluctuations. Based on it we can try to reconstitute the principle milestones in the scientists’ life: the first 5 years correspond roughly to the PhD. studies during which one typically produces a certain (not necessarily high) number of publications. The next 5 – 10 years (2nd period) are of great importance to those who stay in research. In that time authors are evaluated on the international scale and their academic position depends heavily on their productivity. The later stages correspond to the scientific maturity when scientific output stabilizes on average.

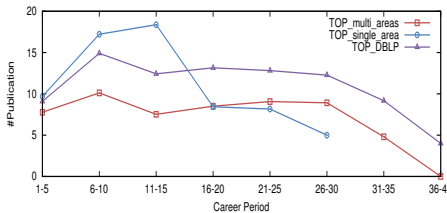


Fig. 2. Author productivity within the different periods of career

With respect to the publication rate values, they are much higher for the single-area authors during the spike periods. There is no additional evidence that would help to explain this phenomenon. We might hypothesize that by working in one field only it is easier to get more papers published, since the author knows better the research criteria of his community.

To analyze the author - publication distribution over venues we calculate for each author $a_i \in \text{TOP}$ dataset the percentage of his publications in the top-ranked conferences relative to all his publications recorded in DBLP. Next we combine the results into the 10%-intervals and match them against the corresponding percentage of authors.

The results are shown on the chart 3a of the Figure 3.

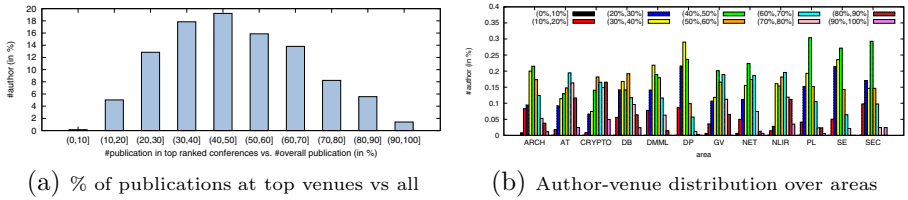


Fig. 3. Authors - venue productivity distribution

It turns out that only about 1.5% of authors in the TOP dataset publish exclusively or mostly at the top-ranked venues. Typically the top-ranked conference publications constitute from 30% to 60% of the author’s conference production. It suggests that the majority of researchers appears in the mixed set of venues.

To look closer at the publication distribution over venues in the topical sets we first assign each author $a_i \in \text{TOP}$ dataset to the area he contributes at most (frequency based majority voting), and perform the same computation as before³.

The chart 3b of the Figure 3 presents the results. Notice that majority of areas are dominated by people who publish between 40 – 50% of their publications in the top ranked conferences, and in DP and DMML the prevailing range is 30 – 40%. These values confirm the general tendency of publishing in the mixed set of venues. On the contrary, authors from DB, CRYPTO, AT and NLIR show more adherence to the top-ranked venues as proportion of researchers who publish 50 – 70% of papers at top-ranked conferences outranks the other categories.

4 Scientific Community Analysis

The previous section dealt with the author characteristic with respect to DBLP and the research areas defined in Section 2. In this section we take a closer look at the areas themselves and investigate them in terms of the *publication growth rate*,

³ CBIO and WWW are not considered as the resulting sets of authors are too small to produce consistent results.

and *population stability*. Selection of the evaluation criteria is not random. We believe that it may help to highlight the peculiarities of the individual domains and compare them to each other. We apply the same set of features to the subset of the non-top ranked conferences and eventually find out the differences between the top and non-top venues.

4.1 Publication Growth Rate

Publication growth rate provides an evidence for the area “well-being” and sheds light on how much interest there is in it at the given moment. It is a dynamic measure that traces yearly changes in the area productivity. We distinguish between the *relative* and *absolute* growth rates and focus on the latter one bellow.

The *absolute growth rate* $AbsGr_{A_i,y}$ of an area A_i in year y is a ratio of publications in A_i within two consecutive years y_i and y_{i-1} such that $AbsGr_{A_i,y} = \frac{Publ_{A_i,y}}{Publ_{A_i,y-1}}$. We have calculated the values for all areas and found that except for the fluctuations corresponding typically to the beginning years, the fields differ considerably from each other. For example, Computer Architecture (ARCH) and Computer Networks (NET) have stabilized at early 90s, their absolute growths rate values oscillate around 1 ± 0.1 . On the contrary, Natural Language Processing and Information Retrieval (NLIR) productivity may vary three times as much from year to year, up to nowadays. Such a diversity could probably result from within-venue conventions that define the number of yearly accepted papers. We therefore compare the conferences in our TOP and NONTOP data sets with regard to the absolute publication growth rate. It turns out to be systematically higher in the non-top conferences. We can translate this result in terms of *publication acceptance rates* (information that is typically not present in the bibliographic databases though it is one of the important parameters for conference evaluation [13,10]), and conclude that they are lower for the top venues.

4.2 Population Stability

In this section we concentrate on the mechanism that influence researcher dynamics. In the context of this section, the large *communities* corresponding to the research areas are decomposed into the conferences each of which is understood as an individual community.

In [1] it has been pointed out that the membership in a community may be influenced by fact of having “friends” in that community. Thus some researchers are more likely to submit their paper to a conference if they have previously coauthored with someone who had already published over there. We take on this approach and investigate whether this property holds equally in different areas and venues.

Due to the space consideration we only discuss some of the most interesting results. Detailed tables can be found in the full version of the paper.

In the TOP set, all venues in AT and CRYPTO prove stable and moreover are the most stable venues in the whole TOP set. They are characterized by

low percentage of newcomers, pure newcomers⁴, and leavers, compared to the average values across the whole TOP set. Note that fraction of pure newcomers is an important parameter as it sheds light on how “friendship” phenomenon affects the inflow of the new authors: the higher the fraction is, the smaller is the friendship influence. We have found that AT and CRYPTO are friendship driven as about 50% of new authors joining venues have co-authored with authors who had already published over there.

Contrarily to the two fields above, WWW conferences are the most dynamic ones, featured by the high values for the newcomers, pure newcomers, and leavers’ fractions. Friendship does not seem to alter the influx of new authors as the pure newcomers typically count for $\approx 60 - 80\%$ of all the newcomers.

The key observation concerning the NONTOP set of venues, is that all of them irrespective of time span (which ranges from 17 to 3 years) and domain, are very dynamic. Typically the newcomers constitute about 75 – 85% of all authors, and the average value of the pure newcomers is about 75% which suggests that the friendship influence on the decision to join a venue is rather negligible. The turnover of authors is also remarkable since the fraction of leavers is often comparable to that of Newcomers and constitutes up to 88% of all the authors. As such, population stability might be considered as a candidate feature that helps to distinguish between the top and non-top venues.

5 Conclusions and Future Work

In this paper we have analyzed computer science communities in different settings. We performed statistical analysis of authors, and found that the DBLP community is dominated by the short-time researchers whose career does not exceed 5 years. We have also discovered that experienced scientists from the top-ranked venues tend to join multiple research communities and produce the highest number of publications between the 5th and 10th years of their career. Typically they publish in a mixture of top and non-top ranked venues.

We have also compared communities from 14 research areas of computer science and performed the between-area comparison in terms of publication growth rate, and population stability. In addition, we applied the same criteria to the comparison between top and non-top ranked conferences and discovered that the publication growth rate and population stability could be among the features that help to separate the two sets.

In this approach we have manually divided the broad area of computer science into 14 topics. In the future we plan to substitute this rather ad hoc approach by applying a machine learning technique such as *Latent Dirichlet Allocation* [3] for both - topic classification and learning the best number of topics into which the given data can be divided. By doing this we will avoid the subjectivity of manual classification. We also plan to elaborate on the set of features that could be used

⁴ *Pure newcomer* is an author who did neither publish at c_k before year y , nor has coauthors $\in c_k$.

for efficient comparison and eventually automatic ranking of venues. Besides we plan to extend the notion of “venue” to incorporate journals into analysis.

Acknowledgment. We would like to thank Prof. Schommer for his valuable comments and suggestions.

References

1. Backstrom, L., Huttenlocher, D.P., Kleinberg, J.M., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: KDD (2006)
2. Bird, C., et al.: Structure and dynamics of research collaboration in computer science. In: Jonker, W., Petković, M. (eds.) SDM 2009. LNCS, vol. 5776, pp. 826–827. Springer, Heidelberg (2009)
3. Blei, D., Ng, A., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, 993–1022 (2003)
4. Elmacioglu E., Lee, D.: On six degrees of separation in DBLP - DB and more. SIGMOD Record (2005)
5. Giles, C.L., Council, I.G.: Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Pnas* 101, 51 (2004)
6. Huang, J., Zhuang, Z., Li, J., Giles, C.L.: Collaboration over time: Characterizing and Modeling Network Evolution. In: WSDM (2008)
7. Newman, M.E.J., Barabasi, A.L., Watts, D.J.: The structure and dynamics of networks. Addison-Wesley Publishing Company, Reading (2006)
8. Sidiropoulos, A., Manolopoulos, Y.: A new perspective to automatically rank scientific conferences using digital libraries. *Inf. Process. Manage.* 41(2), 289–312 (2005)
9. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. *Nature*, 440–442 (1998)
10. Yan, S., Lee, D.: Toward alternative measures for ranking venues: a case of database research community. In: JCDL, pp. 235–244 (2007)
11. Zaïane, O., Chen, J., Goebel, R.: Mining research communities in bibliographical data. In: WebKDD/SNA-KDD, pp. 59–76 (2007)
12. Zhou, D., Ji, X., Zha, H., Giles, C.L.: Topic evolution and social interactions: how authors effect research. In: CIKM, pp. 248–257 (2006)
13. Zhuang, Z., Elmacioglu, E., Lee, D., Giles, C.L.: Measuring conference quality by mining program committee characteristics. In: JCDL, pp. 225–234 (2007)

Citation Graph Based Ranking in Invenio

Ludmila Marian¹, Jean-Yves Le Meur¹, Martin Rajman², and Martin Vesely²

¹ European Organization for Nuclear Research
CH-1211 Geneve 23, Switzerland

{ludmila.marian, jean-yves.le.meur}@cern.ch

² Ecole Polytechnique Fédérale de Lausanne
LIA, Station 14, CH-1015 Lausanne, Switzerland
{martin.rajman, martin.vesely}@epfl.ch

Abstract. Invenio is the web-based integrated digital library system developed at CERN. Within this framework, we present four types of ranking models based on the citation graph that complement the simple approach based on citation counts: time-dependent citation counts, a relevancy ranking which extends the PageRank model, a time-dependent ranking which combines the freshness of citations with PageRank and a ranking that takes into consideration the external citations. We present our analysis and results obtained on two main data sets: Inspire and CERN Document Server. Our main contributions are: (i) a study of the currently available ranking methods based on the citation graph; (ii) the development of new ranking methods that correct some of the identified limitations of the current methods such as treating all citations of equal importance, not taking time into account or considering the citation graph complete; (iii) a detailed study of the key parameters for these ranking methods.

Keywords: CDS, Invenio, Inspire, citation graph, PageRank, external citations, time decay.

1 Introduction

Invenio is the integrated digital library system developed at CERN [4], suitable for middle-to-large scale digital repositories (100K-10M records). It is a suite of applications which provides the framework and tools for building and managing an autonomous digital library server. Besides being used to run the CERN Document Server (which is ranked 4th in the *Webometrics Top 400* institutional repositories [3]), Invenio has also been chosen by several other important institutions and projects. Among them, the recently launched INSPIRE service, that is meant to become the reference repository for High Energy Physics documents. At CERN, Invenio manages over 500 collections of data, consisting of over 1M bibliographic records [1].

In the setting of this framework, our goal is to develop robust citation ranking methods. We start our analysis from existing citation ranking methods, studying their strengths and their weaknesses. We do an in-depth analysis of the set

of parameters that influence the outcome. We develop novel citation ranking methods in order to overcome the identified drawbacks of the existing ones.

We use as a baseline the Citation Count. This method fails in capturing the differences between citations in importance as well as in publication date. In order to take into account the importance of the citations we study link-based ranking methods (Section 4.2). The idea of applying link-based methods to the citations graph is not new (Section 2). We found it relevant to re-evaluate the outcome as well as the parameter analysis in the context of our data sets, and using different metrics. By doing this, we discovered several drawbacks generated by different properties of the citation graph (connectivity, completeness and correctness). We correct these drawbacks by developing a novel link-based ranking that accounts for external citations (Section 4.4). In order to take into account the publication date of each citation, i.e. the “freshness” of the citations, we study time-dependent citation ranking methods (Sections 4.1 and 4.3). Although the decayed time factor was also introduced previously in the literature, our contribution is firstly, applying this method on top of citation counts, introducing the notion of decayed citation counts, and doing an in-depth analysis of the stability of the rankings with respect to the decay factor, and secondly, analyzing time-decayed link-based ranking in the context of our data sets. This led to the discovery of cycle-induced anomalies, that proven this method unsuited for time inconsistent data sets. These methods bring major improvements over the citation count baseline: by considering the importance of the citations, we can identify modestly cited publications that have a high scientific impact on the research community; on the other hand, by taking into consideration the publication date of each citations, i.e. the “freshness” of the citations, we can identify currently relevant publications, or better said, the “hot trends” of a specific domain that would have not been identified by the citation count method.

2 Related Work

In this section we review some of the work that has been conducted in the domains of citation analysis and ranking scientific publications.

In different cases, the citation count is not able to fully capture the importance of a publication, mainly due to the fact that it treats all the citations equally, disregarding their differences in importance and also their creation date. In order to overcome these drawbacks, several studies had been done. P. Chen et al. in [8] apply the Google’s PageRank algorithm (proposed by S. Brin, L. Page in [13]) on the citation graph to assess the relative importance of all publications in the Physical Review family of journals from 1893-2003. They prove with different examples that applying PageRank is better at finding important publications than the simple citation count. They also argue about using a different damping factor than the one used in the original PageRank algorithm. The authors extended their work in [5] by introducing a new algorithm, CiteRank, a modification of PageRank, that also accounts for the date of the citations by distributing the random surfers exponentially with age, in favor of more recent publications. By

this, they try to model the behavior of researchers in search for new information. They test their model on all American Physical Society publications and the set of high-energy physics theory (hep-th) preprints. They find the parameters for their model by trying to maximize the correlation between the CiteRank output and the download history. Also, N. Ma et al., in [12] apply PageRank on the citation graph in order to evaluate the research influence of several countries in the Biochemistry and Molecular Biology fields.

There has been some research activity also in the area of “temporal link analysis”, mostly done on WWW pages. In [6] the authors present several aspects and uses of the time dimension in the context of Web IR. K. Berberich et al. [7] argue that the freshness of web content and link structure is a factor that needs to be taken into account in link analysis when computing the importance of a page. They provide a time-aware ranking method and through experiments they conclude on the improvements brought by it to the quality of ranking web pages. They test their approach on the DBLP data set but with the scope of ranking researchers rather than publications.

The task of ranking scientific documents is a complex one and it should not depend only on the citation graph information. In the Invenio framework, there has been significant work done in trying to aggregate different metrics (i.e. the download frequency, the publication date) in order to create a better suited ranking for scientific documents [11], [10].

3 Experimental Framework

The experiments were conducted on two data sets of bibliographic data (not completely disjoint): Inspire (<http://hep-inspire.net>) containing 500,000 High Energy Physics (HEP) documents and CERN Document Server (<http://cdsweb.cern.ch>) containing 200,000 CERN documents.

We analyzed three important characteristics of the citation graphs extracted from these data sets: *graph connectivity* (i.e. the number of publications that have no citations, the number of publications that have no references), *graph completeness* (i.e. the number of publications missing from the data set) and *graph correctness* (i.e. if the graph allows cycles). The first two characteristics will be discussed per data set basis, while the third, since it is common for both citation graphs, will be discussed separately.

Inspire Data Set. Inspire is a new High Energy Physics information system which will integrate present databases and repositories to host the entire corpus of the HEP literature and become the reference HEP scientific information platform worldwide. It is a common project between CERN, DESY, FERMILAB and SLAC [2].

The Inspire data set contains almost half a million publications, with a total number of 8 million citations. Approximately 25% of the documents are not cited by any other document in the system, while approximately 16% of the documents have no references. On average, a paper is missing 9 references. We

computed this number as the difference between the total number of references displayed for a record and the number of references existing in the database. This 9 missing references/paper, compared with the average number of references that are in the system, 20 references/paper, tell us that although we do not have a complete citation graph, having more than 50% of the references is still better than expected. Also, Inspire is a human edited repository, meaning that the citation extraction is validated by an authorized person.

CERN Document Server Data Set. CDS contains the CERN collection of publications [1]. Out of more than 900,000 bibliographic records indexed by CDS we sampled a subset of 200,000 documents with 1,4 million citations. Approximately 20% of these documents are not cited by any other document in the system while 35% of the documents have no references. On average, each document is missing 28 out of 37 references. One reason for this low number of available references is that currently CDS is using an automated references extractor [9]. Since the future of bibliographic repositories is the automation of the data extraction, one must consider these drawbacks in the development and analysis of the ranking methods based on the citation graph. So, since the Inspire data set generates a better citation graph than the CDS data set, in terms of completeness, we will mainly discuss our results on the Inspire data set, but we will also present solutions for less dense data sets.

Data Correctness. While the intuition is that the citation graph is a directed acyclic graph (DAG), we discovered that this is not true. Since the system contains preprints (drafts of scientific papers that have not yet been published in a peer-reviewed scientific journal) as well as published papers and conference proceedings, it might happen in some cases that future work is cited. On top of this, there are also some cases where a paper is citing itself. We try to eliminate these last types of anomalies as often as possible. Still, the first class of problems is harder to permanently eliminate, and even though theoretically impossible, the “future work” citation is sometimes legitime. For these reasons, we build our algorithms on top of a general directed graph and not on top of DAG.

4 Ranking Methods

In this section we study four types of citation ranking algorithms with respect to the baseline algorithm, Citation Count. All algorithms and parameters have been studied in the context of both data sets. We chose to present the results obtained on the Inspire database, since both connectivity and completeness parameters were higher in this case, thus facilitating the evaluation of the outcome. The only exception is the link-based ranking with external citations (Subsection 4.4), developed in particular for data sets with low completeness of graph (in our case, the CDS data set).

Our goal is to develop robust ranking methods based on the citation graph. In order to achieve this, we start from the citation count method, which we consider

the baseline algorithm. We gradually add features and study both their positive and their negative impact on the final outcome. The result is four ranking methods, each suited for different types of publication discovery.

We first study the effect of time in the citation graph by applying a time-decay factor to the citation counts. In this context, we study the rank stability with respect to various settings of the time parameter. Since this method does not take into consideration the importance of different citations, we continue our analysis with a link-based ranking. Here we study the correlation between the damping factor and the bias towards older publications. In this case, our goal is to retrieve publications that are and have been of great interest for the community, although, they are modestly cited. In order to take into consideration both the age of citations and their importance, we combine the decay factor with the link-based ranking. The idea behind this method is that it is able to retrieve modestly cited papers that are at the present time of interest for their community. Unfortunately, this method suffers from cycle-induced anomalies. Last, but not the least, in order to overcome the bias of PageRank to incomplete citation graphs we introduce a novel link-based method.

4.1 Time-Dependent Citation Count

To overcome the fact that the Citation Count method does not take into account the time dynamics of the citation graph we introduce the notion of *time-dependent citation counts*. In this context, the weight of a publication i is defined as: $weight_i = \sum_{j, j \rightarrow i} e^{-w(t_{present} - t_j)}$ where $t_{present}$ is the present time and t_j is the publication date for document j^{th} .

Furthermore, this introduces the time decay parameter ($w \in (0, 1]$), which quantifies the notions of “new” and “old” citations (i.e. publications with ages less than the time decay parameter would be considered “new”; publications with ages larger than the time decay parameter would be considered “old”). The larger the time decay parameter is, the faster we “forget” old citations.

Results. Since the time decay factor, w , is the only quantifier for the “freshness” of the results, we analyzed its impact on the stability of the final rankings and also on the stability of certain ranges of ranks.

In order to find out if the adding a time decay has a global impact on the ranking (i.e. the tail is promoted to the head and the other way around) or if it is rather local (i.e. there are certain windows in the ranking where there is some reshuffling) we measured the “*locality of changes*”.

Let us consider s as being the *stability factor*: $s = \frac{\{d | rank_d(t), rank_d \in window\}}{windowSize}$ where $rank_d(t)$, $rank_d$ are the ranks of publication d , the first when using a time-dependent ranking method and the last when using the non-decayed ranking method.

Using the stability factor we want to determine what windows of the ranking are suffering the most from different time decay parameters. For this we are building dynamic windows as follows: we are splitting the rank range by consecutive powers of 2, until we either reach a rank window of size less than 100 or the stability factor goes below a certain minimum threshold (0.3 in our experiments).

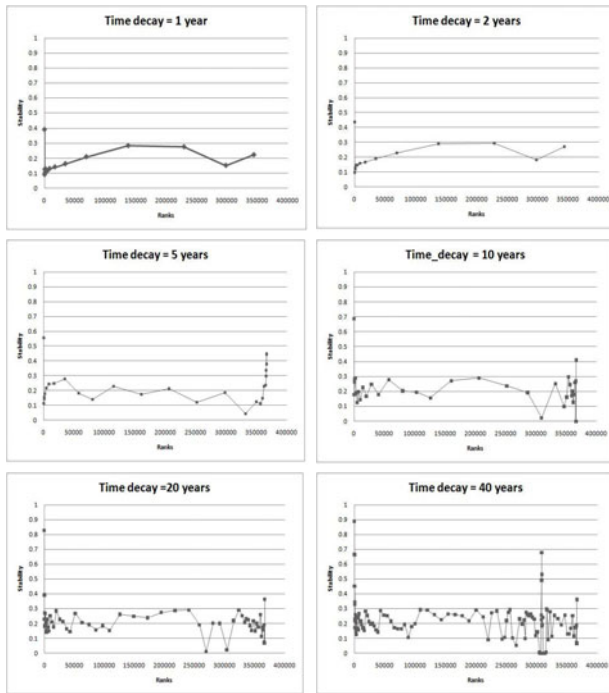


Fig. 1. Stability of Time-dependent Citation Count

We should mention that we remove the publications with 0 citations (125k documents), since their weight and rank will not be influenced by any ranking method. We constructed a chart for each value of the time decay factor (1 year, 2 years, 5 years, 10 years, 20 years, 40 years) (Figure 1). The interpretation of these charts is that whenever we have a zone with a lot of activity (a lot of points), that zone is quite stable at a high level and needs to be broken into small intervals to reach the instability threshold. On the other hand, when we have a zone with low activity, that means that the stability of the corresponding window is low also at a high level, so if we would split it in smaller windows, the stability will drop even lower than the threshold. From the Figure 1 we observe that the head of the ranks is usually more stable than the rest. Also, even with such a large time decay as 40 years, the ranks are still reshuffled, but in small windows.

We also analyzed the effects of different values of the time decay factor on the number of publications promoted/demoted. We discovered that that the *time-depending ranking methods are promoting more publications than demoting*. Secondly, and as a consequence of the first observation, the *time-depending ranking methods are demoting strongly than promoting*.

Based on this analysis one can choose between either having a strong time decay, which will boost really new publications, or having a weaker time decay,

which will still boost publications with newer citations, but will also take into account old citations.

Adding even a weak time decay factor, the time-dependent ranking can still differentiate between an old publication that acquired a large number of citations over a long period of time, and a new publication, that, although important for the scientific community, did not have enough time to acquire as many citations as the old one, in the favor of the latter. Still, this method inherits one major shortcoming from the Citation Count method, i.e. it does not take into consideration the different importance of each citation. To overcome this, we developed the Time-dependent Link-based Ranking as a combination of Time-dependent Citation Counts and Link-based Ranking (Subsection 4.3).

4.2 Link-Based Ranking: PageRank on the Citation Graph

The PageRank algorithm [13] is based on a random surfer model, and may be viewed as a stationary distribution of a Markov chain.

The PageRank model assigns weight to documents proportional with the importance of the documents that link to them:

$$PR(p_i) = \frac{1-d}{n} + d \sum_{j:p_j \rightarrow p_i} \frac{PR(p_j)}{deg(j)} \quad (1)$$

where $PR(p_i)$ is the PageRank score of paper i and $deg(j)$ is the out-degree of node j (i.e. total number of documents cited by paper j). d is called *damping factor* and in the literature concerning the web graph it usually has values in $[0.85, 1)$. It is a free parameter that controls the performance of the PageRank algorithm, preventing the overweighing of older publications. This ranking models the behavior of a user moving from paper to paper in the document collection [8]. At each moment in time the user can either follow a randomly chosen reference from the current document, with the probability d , or he can restart the search, from a uniformly randomly chosen publication with a probability of $1-d$. For the WWW it is considered that on average, the users follow 6 continuous links, until they get bored and restart the search. In [8] the authors consider that a researcher will only follow on average 2 links on the citation graph, until the search is restarted. This is why they propose a damping factor of 0.5. In order to verify their hypothesis, we tested three different values for the damping factor: 0.50, 0.70, 0.85.

Results. The calculation of the Spearman's rank correlation coefficients generated with the three chosen values for the damping factor, showed us that, at the global scale, the differences between rankings are almost undetectable (the lowest correlation, with a value of 0.996 was between $d=0.50$ and $d=0.85$). So in order to choose the best d we have to dig deeper. For this, we looked at the distribution of the ranks over the time. Since we know that a higher damping factor is boosting old papers rather than new ones, we are interested to see if we can detect this kind of behavior also for our data. For this, we plotted the distribution in time for the Top 100 papers ranked with PageRank. The results

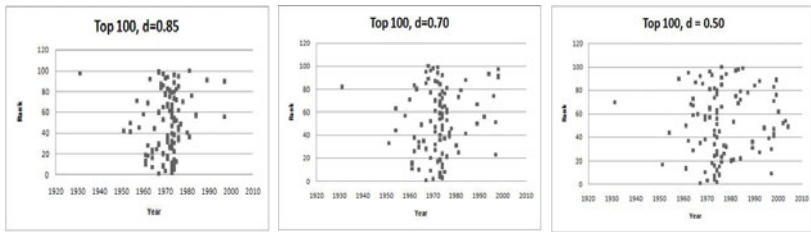


Fig. 2. Time distribution for the top 100 publication, ranked with PageRank

Table 1. Top 10 publication by PageRank, when damping factor = 0.50 (CC = Citation Count, RCC = Rank by Citation Count)

CC	Publication	Rank	RCC
6565	A Model of Leptons: Weinberg, Steven (1967)	1	1
3023	Confinement of Quarks: Wilson, Kenneth G., (1974)	2	26
3671	Weak Interactions with Lepton-Hadron Symmetry: Glashow, S.L., (1970)	3	9
5351	CP Violation in the Renormalizable Theory of Weak Interaction: Kobayashi, Makoto, (1973)	4	2
2379	Ultraviolet Behavior of Nonabelian Gauge Theories: Gross, D.J., (1973)	5	44
2472	Radiative Corrections as the Origin of Spontaneous Symmetry Breaking: Coleman, Sidney R., (1973)	6	40
2390	Reliable Perturbative Results for Strong Interactions?: Politzer, H.David, (1973)	7	43
1978	Pseudoparticle Solutions of the Yang-Mills Equations: Belavin, A.A., (1975)	8	56
3556	Maps of dust IR emission for use in estimation of reddening and CMBR foregrounds: Schlegel, David J., (1997)	9	13
2332	Axial vector vertex in spinor electrodynamics: Adler, Stephen L., (1969)	10	47

are displayed in Figure 2. Indeed, we can see that for a damping factor of 0.5, the age of the top 100 papers decreases. If for a damping factor of 0.85 we have a large concentration of top papers in the 1970-1980 period, when decreasing the d , we see a shifting of the top papers towards the 1990-2000 period. Since we wish to have a ranking that is not biased towards older publications, we also conclude that a value of 0.5 for the damping factor is better suited.

The main advantage of this ranking method is that it weighs each publication based on the importance of its citations. In this way, the quality is preferred over the quantity. We can say that it associates to each publication an “all-time achievement” rank (Table 1).

4.3 Time-Dependent Link-Based Ranking

The idea of the Time-dependent Link-based Ranking method is to distribute the random surfers exponentially with age, in favor of more recent publications. Every researcher, independently, is assumed to start his/her search from a recent paper or review and to subsequently follow a chain of citations until satisfied. In this way the effect of a recent citation to a paper is greater than that of an older citation to the same paper. This method was also presented in [5].

We consider the weight of each publication as being inversely proportional with its age: the younger the publication is, the more its citations will value. In this case, the initial probability of selecting the i^{th} paper in a citation graph will be given by: $p_i = e^{-w(t-t_i)}$, where t is the present time, t_i is the publication date for document i^{th} and w is what we call the *time decay parameter*. Adding the time decay to equation (1), we obtain:

$$PR(i, t) = \sum_{x=1}^n \left(\frac{1-d}{n} \times p_x(t) \right) + d \sum_{j,j \rightarrow i} \left(\frac{PR(j)}{deg(j)} \times p_j(t) \right)$$

$p_x(t)$ is the probability of initial selecting the x^{th} node in the citation graph.

Results. Analyzing the ranking results, we discovered in Top 100 cases of older publications, with a modest number of citations, which, due to the fact that they acquired some of these citations recently, are ranked higher compared with the PageRank score, and so, they are easier to be discovered by the researchers. This is exactly the outcome we were hoping to see. Unfortunately, we also discovered some anomalies (Table 2).

The two publications presented in Table 2 have less than 20 citations, and thus, are ranked really low with the Citation Count ranking method. How is it possible to be so highly ranked with the new ranking method? Further investigations showed that the problem comes from the fact that these two papers are citing each other, and thus, are part of a cycle. Because of this and of the link-based ranking which iteratively propagates the weight in the graph, when a strong time decay factor is used (in our case, a 5 year time decay), the newly published documents that are part of a cycle accumulate artificial weight. Unfortunately, this makes the time-dependent link-based ranking method unsuitable for data sets that allow cycles. As discussed previously, even the bibliographic data sets can allow cycles due to certain inconsistencies in the publication dates

Table 2. Snapshot from Top 100 publications by Time-dependent PageRank (CC = Citation Count)

Citations	Publication	Rank	Rank by CC
19	Gauge symmetry and supersymmetry of multiple M2-branes: Bagger, Jonathan (2007)	31	90786
18	Comments on multiple M2-branes: Bagger, Jonathan (2007)	32	94900

or in the listing of references. Since some of the publications are not dated, the identification/removal of the cycles is almost impossible due to the computational overhead. Because of this and of the link-based ranking which iteratively propagates the weight in the graph, when a strong time decay factor is used, the newly published documents that are part of a cycle accumulate artificial weight. Thus, this method is not suited for data sets that allow cycles.

4.4 Link-Based Ranking with External Citations

As we saw in Section 3, the Inspire data set is missing on average 9 out of 30 references per paper while the CDS data set is missing on average 28 out of 37 references per paper. While for the Inspire data, these missing links represent just a small percentage, for the CDS data they represent almost 75%. In the context of applying the PageRank algorithm, this means that instead of distributing the weight to 37 references, a node is distributing its weight only to 9. This further means that these 9 papers receive much more weight than expected. So, we end up with a phenomena of “artificial inflation of weights”.

For fixing this error we developed a new ranking method that accounts for the external citations. This new method assumes the existence of an “external authority” that accumulates weight from all the nodes in our graph, proportionally with the missing citations, and also feeds back into the network a certain percentage of its weight. With this method, we assure the correct propagation of the weight through the network.

The “external authority” (EA) node is controlled by two parameters, α and β . Each publication will contribute to the EA’s weight with $\frac{\beta \times \max\{1, ext_i\}}{\beta \times \max\{1, ext_i\} + int_i}$, where ext_i is the number of external citations for publication i , and int_i is the number of internal citations. On the other hand, EA contributes to all publications with $\frac{\alpha}{n}$ weight, where n is the total number of publications in the repository. Intuitively, α quantifies how much of the external weight is re-injected into the network and β represents the fraction between an external citation and an internal one. We consider that, if a publication is not in the data set, it means that it values less for the repository than the ones already inserted in the database.

Results. In order to analyze how α and β influence the final outcome of the ranking we calculated the Spearman Correlation Coefficient (SCC) between our new ranking method with different settings of α and β (between 0 and 1 with 0.1 step), and the PageRank, for the CDS data set.

Table 3 presents the aggregated results after 200 experiments (for each $\alpha, \beta \in (0, 1)$, with a step of 0.1). Our experimental analysis showed that α only influences the rate of convergence of the iterative algorithm (with the best convergence rate obtained for $\alpha = 0.5$) and has little impact on the general reordering while β is the one that makes a difference in the outcome of the ranking method. For $\beta \in [0.1, 0.5)$ the outcome of the new ranking method is highly correlated with the PageRank results, and less correlated with the Citation Count results. On the other hand, for $\beta \in [0.5, 1]$ the correlation with the PageRank method drops, while the correlation with the Citation Count remains approximately constant. We advise for the use of a β lower than 0.5 since in this case the results

Table 3. Spearman Correlation Coefficient between PageRank/Citation Count and Ranking with External Citations (The SCC between the PageRank and the Citation Count is 0.81)

α	β	SCC with PageRank	SCC with Citation Count
$\alpha \in (0, 1)$	$\beta = 0.1$	0.97	0.89
$\alpha \in (0, 1)$	$\beta = 0.2$	0.94	0.91
$\alpha \in (0, 1)$	$\beta = 0.3$	0.92	0.92
$\alpha \in (0, 1)$	$\beta = 0.4$	0.91	0.92
$\alpha \in (0, 1)$	$\beta = 0.5$	0.89	0.93
$\alpha \in (0, 1)$	$\beta = 0.6$	0.88	0.93
$\alpha \in (0, 1)$	$\beta = 0.7$	0.87	0.93
$\alpha \in (0, 1)$	$\beta = 0.8$	0.87	0.93
$\alpha \in (0, 1)$	$\beta = 0.9$	0.86	0.93
$\alpha \in (0, 1)$	$\beta = 1.0$	0.85	0.93

will be less correlated with the citation counts and enough correlated with the PageRank as to assume that the artificial inflation problems are resolved. We believe Link-based Ranking with External Citations to be a better candidate than Citation Count or PageRank for the task of ranking scientific publications because: (i) it inherits from PageRank its ability to take into account the citations with weights representing their importance, and thus, fixing one of the main shortcomings of the Citation Count method; (ii) it further corrects one of PageRank's shortcomings, namely the artificial inflation of some of the weights. In the end, our new ranking method is enough correlated with the PageRank method as to assume that it inherits its usefulness and in the same time it corrects its shortcomings.

5 Conclusions

The Citation Count is a very popular measure of the impact of a scientific publication. Unfortunately, it has two main disadvantages: it gives all the citations the same importance and it does not take into account time. These drawbacks motivated our study of alternative approaches: Time-dependent Ranking methods and Link-based Ranking methods. The time-dependent ranking methods were developed to take into account the time dynamics of the citation graph. More precisely, we first introduced time-dependent citation counts, taking into consideration the lifetime of the citations. Finally, we combined the link-based ranking with the time-dependent citation counts, creating the Time-dependent Link-based Ranking. Unfortunately, this algorithm is not well suited for the citation graphs that are not DAG, due to the fact that it tends to overweight the young publications that are part of a cycle. The link-based ranking methods were developed to take into account the importance of the citing papers. We started with the PageRank algorithm originally designed for ranking web pages. In order to make it better suited for the bibliographic citation graph, we first modified the setting of the damping factor. Furthermore, we adjusted

the PageRank model by adding an “external authority” node that represents a place holder for all the missing citations. In particular, this additional node prevents some publications from getting artificially boosted simply because of the incompleteness of the citation graph. We believe Link-based Ranking with External Citations to be a better candidate than Citation Count or PageRank for the task of ranking scientific publications.

In terms of future work, we plan to carry out a study on combining the above mentioned ranking methods that are based on citations with other ranking methods that are available in the CDS Invenio software, notably the download counts, word similarity, and reputation measures such as the Hirsch Index.

Acknowledgments. We would like to thank our *CDS Invenio* colleagues for their continuous support during this project. Also, we would like to thank Travis Brooks, Inspire Project coordinator, for fruitful discussions.

References

1. Cds invenio, <http://cds.cern.ch/>
2. Inspire project, <http://inspire.cern.ch>
3. Webometrics, http://repositories.webometrics.info/top400_rep_inst.asp
4. Gracco, M., Le Meur, J.-Y., Robinson, N., Simko, T., Pepe, A., Baron, T., Vesely, M.: Cern document server software: the integrated digital library (2005), <http://doc.cern.ch/archive/electronic/cern/preprints/open/open-2005-018.pdf>
5. Walker, D., et al.: Ranking scientific publications using a simple model of network traffic (2006)
6. Amitay, E., et al.: Trend detection through temporal link analysis. In: J. of the American Society for Information Science and Technology, pp. 1–12 (2004)
7. Berberich, K., et al: T-rank: Time-aware authority ranking. In: WAW, pp. 131–142 (2004)
8. Chen, P., et al.: Finding scientific gems with google. J. Informet. 1, 8–15 (2007)
9. Le Meur, J.-Y., Claivaz, J.-B., Robinson, N.: From fulltext documents to structured citations: Cern’s automated solution (2001)
10. Rajman, M., Vesely, M., Le Meur, J.-Y.: The d-rank project: Aggregating rankings for retrieval of scientific publications in the hep domain (2008)
11. Rajman, M., Vesely, M., Le Meur, J.-Y.: Using bibliographic knowledge for ranking in scientific publication databases, pp. 201–212 (2008)
12. Zhao, Y., Ma, N., Guan, J.: Bringing pagerank to the citation analysis, pp. 800–810 (2007)
13. Page, L., Brin, S.: The anatomy of a large-scale hypertextual web search engine. In: Computer Networks and ISDN Systems, pp. 107–117 (1998)

A Search Log-Based Approach to Evaluation

Junte Zhang¹ and Jaap Kamps^{1,2}

¹ Archives and Information Studies, Faculty of Humanities, University of Amsterdam

² ISLA, Faculty of Science, University of Amsterdam

Abstract. Anyone offering content in a digital library is naturally interested in assessing its performance: how well does my system meet the users' information needs? Standard evaluation benchmarks have been developed in information retrieval that can be used to test retrieval effectiveness. However, these generic benchmarks focus on a single document genre, language, media-type, and searcher stereotype that is radically different from the unique content and user community of a particular digital library. This paper proposes to derive a domain-specific test collection from readily available interaction data in search logs files that captures the domain-specificity of digital libraries. We use as case study an archival institution's complete search log that spans over multiple years, and derive a large-scale test collection. We manually derive a set of topics judged by human experts—based on a set of e-mail reference questions and responses from archivists—and use this for validation. Our main finding is that we can derive a reliable and domain-specific test collection from search log files.

1 Introduction

A digital library (DL) is created for domain-specific collections, whether it be in cultural heritage or about scientific articles. But how good are DLs for disclosing their collections for their particular user groups? Anyone who is interested in DLs probably has already asked this question [8]. A major challenge for DLs is how to evaluate the information retrieval (IR) effectiveness given the domain-specificity of their collections, and how to use this crucial evaluation step to improve a DL.

In IR, the dominant approach to evaluation uses a test collection: a set of documents, a set of search requests (topics), and a set of relevance judgments for each topic (*qrels*). Such test collections are created collaboratively on generic (artificial) set of documents, such as newspaper corpora or Wikipedia, and are useful to study generic aspects of retrieval models. Such test collections provide part of the answers, but fail to address the unique collection and types of search requests of an individual DL. Creating a test collection with this conventional approach, for each DL, is simply too expensive in time and effort.

We propose a log-based approach to IR evaluation of DLs. Nowadays, almost every DL is Web-based, and the interaction between the system and the user is logged in so-called search logs, often hidden deeply away or primarily used to generate descriptive statistics about the general Web traffic of a website. This

includes information that has been entered by the user, and what and where it was clicked, and so on. Is it reasonable to assume that such data can be reused for evaluation? This results in this main research question:

- *Can we use a digital library’s search log to derive a domain-specific test collection?*

The envisioned test collection is tailored to the DL at hand, representative to both its document collection and its search requests. As a test collection, it can be (re)used for comparative testing under the same experimental conditions. Performance is topic-dependent and this avoids comparing over different topic sets. We apply this approach to a particular domain-specific collection of documents, in a special genre, namely archival finding aids for archives of persons, families, and corporations. These archival finding aids are created in electronic form to provide online archival access, using the Encoded Archival Description (EAD, [17]) standard based on Extensible Markup Language (XML, [4]).

The remainder of this paper is structured as follows. Section 2 describes related work. An archival institution’s complete search log that spans over multiple years is used in our experimentation. We derive a domain-specific test collection from a search log in Section 3. We deploy the resulting domain-specific test collection for evaluation using a range of retrieval models in Section 4. In order to validate the log-based evaluation, we construct a set of topics judged by human experts—based on a set of e-mail reference questions and responses from archivists. The results are analyzed and discussed in Section 5. Finally, we conclude with our main findings and discuss pointers to future work in Section 6.

2 Background and Related Work

In this section we discuss three strands of related work: transaction log analysis, IR evaluation, and archival (metadata) retrieval.

2.1 Log Analysis for Information Retrieval

Historically, the analysis of log files started and “evolved out of the desire to monitor the performance of computerized IR systems” [16, p.42]. The focus has been to analyze how systems are used. Besides system monitoring, it can also be conceptualized as a way to unobtrusively observe human behaviors. Studies in a DL setting have been reported in [13], which focused particularly on the queries that users entered in the system, with the proposition that the analysis can be used to finetune a system for a specific target group of users, but it did not investigate the IR effectiveness.

Research on log analysis in library and information science preceded the research in the World Wide Web, where the latter zooms into IR by analyzing search engines. An overview on search log analysis for Web searching, and a methodology, is presented in [11], which shows that literature on log analysis for Web-searching is abundant. The logs can be analyzed to better understand

how users search on the Web effectively. An example is the paper of [23], which describes a study about search logs, where the search behavior of advanced and non-advanced users is analyzed by testing the effects of query syntax with query operators on query-click behavior, browsing behavior, and search success.

There has been substantial interest in using clickthrough data from search logs as a form of implicit feedback [5]. Joachims et al. [12], p.160 conclude that “the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages.” There is active research on building formal models of interaction from logs to infer document relevance [6].

2.2 IR Test Collections

IR evaluation can be traced back to the workprocess of a librarian working with card indexes using library classification schemes [19]. The basic methodology for IR experimentation has been developed in the 1950s with the Cranfield experiments, focusing on retrieval effectiveness by the comparative evaluation of different systems (indexing languages in the 1950, retrieval algorithms nowadays). Much of the experimentation focuses on building a ‘test collection’ consisting of a document collection, a set of topics and judgments on which documents are relevant for each topic [21]. Test collections can be reused by evaluating new or adapted systems or ranking algorithms under the exact same experimental conditions. There exist test collections for a variety of domains. Examples included the Cystic Fibrosis database [20], and WT10g for the Web in general [2]. In the field of Focused Retrieval, the Initiative for the Evaluation of XML Retrieval (INEX) constructed test collections from XML files [14].

2.3 Evaluating Archival Metadata Retrieval

Published research that empirically or experimentally deals with the evaluation of archival metadata retrieval is scant [10]. Experiments that specifically examine the retrieval performance potential of archival finding aids in specifically EAD is almost non-existent, despite the emergence of EAD in 1997 [17] and its increasing adoption and popular use in archives.

The first study in the archival field that empirically tested different subject retrieval methods was Lytle [15]. Subsequently, there were a few studies that tested the effects of some external context knowledge on retrieval, such as controlled vocabulary terms [18] or document-collection granularity [10]. The retrieval of online archival finding aids (not in EAD) have been examined in the study of [7] by counting the number of finding aids returned by search engines using different types of query reformulations, i.e. keyword, phrase, and Boolean searches using the topical subject and names headings as queries. The retrieval experiments of [22] on finding aids as full text HTML documents on the World Wide Web pointed to the effectiveness of phrases for the retrieval of finding aids (not in EAD) in six IR systems. The only study so far that focused on the use of EAD on the XML element level was [24], which tested the ranking based on relevance and archival context.

3 From Search Log to Test Collection

In this section, we study how to derive a test collection from a search log. We perform a case study of an archival institution, and use its search log to create domain-specific test collections, tailored to the collection and users.

3.1 Search Log Files and Document Collection

We have obtained the search logs of the National Archives of the Netherlands (NA-NL). The history preserved at this institution goes back to more than 1,000 of years, preserved in archives which stretch more than 93 kilometers or 58 miles. It also includes maps, drawings, and photos—much of it is published on the NA-NL website (www.nationaalarchief.nl). The website provides access by offering a search engine, which includes searching in archival finding aids compiled in EAD [17], image repositories, and separate topic-specific databases.

The logs were 91.1 GB in size, with 39,818,981 unique IP-addresses, and collected from 2004 to a part of 2009 on a Microsoft IIS server. This illustrates that the NA-NL attracts high traffic. The information contained in the search logs were recorded from 2004-2006 in the *Common Logfile Format* (CLF), and from 2007 to 2009 in the *W3C Extended Logfile Format* (ELF). The information in the CLF format included a date, a timestamp of a hit, unique identifier for the user, the URL of the link that was visited, the query string, and a browser identifier. In the ELF format, it also included a referral, and transactions were recorded in detail within each second.

In our experiments, we focus on clickthrough data of online archival finding aids in EAD, where each click contains the filename of a result and a corresponding query. The reason is that we have also obtained these matching EAD files for analysis and further experimentation. Each EAD file describes the contents of an entire archival collection. We use 4,885 EAD files in XML—651 MB of data obtained and mostly written in the Dutch language—from the National Archives of the Netherlands, which were also found in the log files. The mean length of the text-only content of these files is 40,608 bytes (median = 9,119), the mean count of the number of XML pair tags is 2,334 (median = 540), thus some of the archival finding aids are exceptionally long in content and complexly and deeply structured in XML.

3.2 Information Extraction from Logs

A DL's search log contains both searching and browsing behavior, with complete sessions starting from an initial query. Given the massive size of the log, we pre-processed it by extracting the clickthrough data that consist of the subset of clicks to EAD URLs. The query string, clicked URL, and the IDs of the user are extracted. It is further processed by aggregating the clicks for each query in a session and keeping track of the count. We define a session as a subset of n clicks from the same IP address, if and only if the difference between i and $i + 1 < 30$ minutes (or 1,800 seconds), where i is a click. This results eventually in 194,138

Table 1. Example of information in sessions extracted from the log

Query (Topic)	File	Session ID	#
burgerlijke stand suriname	1.05.11.16	504d2bbe246d877bda09856ecc300612.5	28
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.1	6
burgerlijke stand suriname	3.223.06	22fe3a65b0c9223280f2dd576c57a012.35	1
burgerlijke stand suriname	1.05.11.16	2b844140ef7cfd438300da7ec6278de0.147	1
burgerlijke stand suriname	2.05.65.01	3784a93938e29a6aef8f50baa845a6f3.1	1
burgerlijke stand suriname	1.05.11.16	8b21ec51722f3a52cfaf35d320dfac0.3	1
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.2	1
burgerlijke stand suriname	1.05.11.16	9235756a6dbdcffba9179d75108cd220.433	1
burgerlijke stand suriname	3.231.07	3c34072bef0d505467ca9394c392888d.2	1

sessions. Table 1 presents the extracted interaction data on an aggregated level. This is used to derive a test collection.

When we focus on Table 1, we notice that for query “burgerlijke stand suriname” (in English, “registry of births, deaths and marriages suriname”) clicks exist in 9 different sessions, coming from 8 different IPs. There were 28 clicks in one session for EAD file “1.05.11.16,” and the same file was clicked in total 6 different sessions. The same file was re-clicked from an IP address in the next session. Henceforth, the EAD file “1.05.11.16” could be regarded as “relevant.”

Although we regard here “clicked pages” as pseudo-relevant, we make no particular claims on the interpretation of clicks. We make the reasonable assumptions that searchers found these pages of sufficient interest—for whatever reason—to consult them more closely, and that a more effective ranking algorithm will tend to rank such pages higher than those that do not receive clicks. In this paper we are interested in the potential of log-based evaluation, and a relatively naive click model is sufficient for that purpose.

3.3 Types of Test Collections

A subset of the search log files is used, namely the clicks on archival finding aids in EAD, which is rapidly growing in use for archival Web access. We notice that the usage of EAD started to take off in 2006 (19.9MB out of 9.8GB; 0.20%), and this trend in popularity was upward, as it also increased in 2007 (1.5GB out of 31.5GB; 4.8%), and in 2008 (2.8GB out of 41.2GB; 6.8%), and a part of 2009 (304.4MB out of 3.8GB; 7.8%). Hence, the Web traffic of the National Archives of the Netherlands is increasingly consisting of the use of EAD, although the amount of EADs published online has increased as well.

We extract from the search log files in total 50,424 unique topics (after string processing, i.e. squeezing white spaces, conversion to lowercase, removal of punctuation), which have been created by 110,805 unique IP-addresses. There were in total 465,089 clicks with 91,009 unique topic-click pairs. Since the collection consists of 4,885 EAD files, numerous topics matched with these files. Table 2 depicts the 8 most popular topics. The queries have a long-tail distribution, where the majority of the topics were unique queries with only 1 hit. This is also

Table 2. Top 8 most popular used query strings, where the total number of clicks with query is 465,089 with 50,424 unique queries

Position	Query String	Count (%)
1	voc	4,383 (0.94)
2	suriname	4,277 (0.92)
3	knil	2,785 (0.60)
4	knvb	2,506 (0.54)
5	wic	1,891 (0.41)
6	hof	1,633 (0.35)
7	hof van holland	1,567 (0.34)
8	arbeidsdienst	1,541 (0.33)

typical in the archival domain, for example genealogists looking for (unique) family names, and this was also the case in the NA-NL logs.

We derive different test collections from the logs. We use clicks on the file-level in order to evaluate full-text retrieval. The two types of test collections used in our experiments are:

Complete Log Test Collection. The set of 50,424 unique topics, and their corresponding clicks to EAD files, where each and any click is treated as a pseudo-relevance judgment.

Test Collections Based on Agreement. Subsets filtered by the agreement among multiple users on the same clicked documents for a given topic. For agreement 2, we only retain documents clicked by at least two users which restricts it to 4,855 topics.

We test the two types of test collections separately in the next section.

4 Log-Based Evaluation in Action

In this section, we use the log-based test collections to determine the retrieval effectiveness of different ranking methods. We look at both the complete log, as well as on smaller subsets based on agreement. Recall that test collections are used for the comparative evaluation of systems or ranking algorithms, hence we need a number of variant systems in order to show their retrieval effectiveness.

4.1 IR Models and Systems

Our system [25] uses MonetDB with the XQuery front-end Pathfinder [3] and the IR module PF/Tijah [9]. All of our EAD files in XML are indexed into a single main memory XML database without stopword removal, and with the Dutch snowball stemmer. To test the effectiveness of the two types of test collections, we use four retrieval models used in PF/Tijah as independent variables. These are controlled by using the default parameter values, the collection λ is set to 0.15—which we find to be working optimal—and we set the threshold of the ranking for each topic to 100.

BOOL is the Boolean model, where there is no ranking, but a batch retrieval of exact matching results. The query is interpreted as AND over all query terms, and the resulting set is ordered by document id.

LM is standard language modeling without smoothing, which means that all keywords in the query need to appear in the result.

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)} \quad (1)$$

where $n(t, q)$ is the number of times term t is present in query q .

LMS is an extension of the first model by applying smoothing, so that results are also retrieved when at least one of the keywords in the query appears.

$$P(t|d) = (1 - \lambda) \cdot P_{mle}(t|d) + \lambda \cdot P_{mle}(t|C) \quad (2)$$

where $P_{mle}(t|C) = \frac{df_t}{\sum_i df_i}$, df_t is the document frequency of query term t in the collection C .

NLLR is the NLLR or length-normalized logarithmic likelihood ratio, is also based on a language modeling approach. It normalizes the query and produces scores independent of the length of a query.

$$NLLR(d, q) = \sum_{t \in q} P(t|q) \cdot \log \left(\frac{(1 - \lambda) \cdot P(t|d) + \lambda \cdot P(t|C)}{\lambda \cdot P(t|C)} \right) \quad (3)$$

OKAPI is Okapi BM25, which incorporates several more scoring functions to compute a ranking, such as also the document length as evidence.

$$BM25(d, q) = \sum_{t \in q} IDF(t) \cdot \left(\frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \right), \quad (4)$$

where we set $k_1 = 2.0$ and $b = 0.25$. We use $IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$, where N is the total number of documents in the collection, and $n(t)$ is the function that counts the number of documents that contains query term t .

4.2 Complete Log Test Collection

In our evaluation, we use three IR measures, namely Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG). The MRR is a static measure that looks at the rank of the first relevant result for each topic, and the nDCG measure that uses the number of clicks on each result by different results as a form of graded relevance judgment.

When we use all topics for evaluation, and look at all measures, we see in Table 3 that *BOOL* is obviously the worst performing system. We note that the differences among the other systems are modest, but these differences are all significant (1-tailed) using the Paired Samples t-Test on a 1% significance level.

Table 3. System-ranking of runs over all topics

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1808 (5)	0.2493 (4)	0.2548 (3)	0.2591 (2)	0.2631 (1)
MRR	0.2015 (5)	0.2940 (4)	0.2980 (3)	0.3024 (2)	0.3077 (1)
nDCG	0.2659 (5)	0.3289 (4)	0.3547 (3)	0.3605 (2)	0.3652 (1)

Table 4. Distribution (in percentages) of topics over query length for all topics compared to when filtering on agreements, and e-mail references, resulting in N topics

# Tokens	All	Agree 2	Agree 3	Agree 4	E-mail
1	37.66	77.65	80.90	79.45	17.81
2	33.43	15.16	12.62	13.27	19.18
3	16.97	5.33	4.89	5.29	30.14
4	6.68	1.15	0.98	1.23	19.18
N	50,424	4,855	2,147	1,304	73

When looking at the recall over all topics, we see that *BOOL* and *LM* retrieved 48,096 relevant results out of 87,057 (55.25%), *LMS* returned 57,935 of 89,906 (64.44%), *NLLR* had a recall of 65.18%, and *OKAPI* returned most relevant results with 65.69%. It shows that the system using the Okapi model performs best with our Dutch document collection, and that exact matching using *BOOL* and *LM* both do not pay off for the early rank (MRR), and as expected hurts the recall. The recall values can be clarified by the long-tail distribution of query terms, which contains many non-occurring names.

In summary, these results are in line with our expectations, namely that *BOOL* would be the worst-performing system, then *LM*, with *LMS* improving over *LM*, and that the differences among *LMS*, *NLLR*, and *OKAPI* would be modest (but significant). We will validate the relative system ranking against a set of humanly judged topics in the next section, but first we will look at the system ranking induced by smaller subsets of topics based on agreement.

4.3 Test Collection Based on Agreements

The log-based test collection has many more topics than a traditional test collection with 25-200 topics. While having thousands of topics opens us new uses, such as focusing on various breakdowns of the topic set even on relatively rare phenomena, it also presents an efficiency challenge: many DLs crumble under thousands of queries. We take into account the agreement that exists among different searchers. For example, when we pay attention to Table 1, this means that only EAD file “1.05.11.16” is included as a relevance judgement in the test collection, and the rest is discarded. We see in Table 4 that as we increase the threshold of agreement, the number of topics decreases significantly. Take for example notice that in the case that if the agreement is set to 2, the topic set

Table 5. System-ranking of runs over topics with agreement

	Agreement	BOOL	LM	LMS	NLLR	OKAPI
MAP	2	0.1522 (5)	0.3605 (4)	0.3620 (3)	0.3629 (2)	0.3751 (1)
	3	0.1120 (5)	0.3891 (3)	0.3888 (4)	0.3894 (2)	0.3991 (1)
	4	0.1071 (5)	0.3637 (4)	0.3639 (3)	0.3641 (2)	0.3793 (1)
MRR	2	0.1629 (5)	0.4020 (4)	0.4030 (3)	0.4039 (2)	0.4157 (1)
	3	0.1188 (5)	0.4253 (2)	0.4247 (4)	0.4253 (2)	0.4356 (1)
	4	0.1132 (5)	0.3943 (3)	0.3942 (4)	0.3945 (2)	0.4110 (1)
nDCG	2	0.2734 (5)	0.4521 (4)	0.4564 (3)	0.4578 (2)	0.4726 (1)
	3	0.2384 (5)	0.4750 (4)	0.4767 (3)	0.4778 (2)	0.4913 (1)
	4	0.2315 (5)	0.4520 (4)	0.4552 (3)	0.4560 (2)	0.4735 (1)

size decreases to 4,855 from 50,424, and when we set the threshold to 4, only 1,304 topics are left over.

What does this mean for evaluating a system with such a set size? The results of this experiments are presented in Table 5. We focus on the differences of the MAP scores when we take an agreement between two different IPs. The *BOOL* is significantly performing worst, and *OKAPI* is performing the best compared to either *LMS* with a significant improvement of 3.61% ($t(4835) = 5.50$, $p < 0.01$, one-tailed), or similarly an 3.35% significant improvement over *NLLR*. Although the difference between *LM* and *LMS* was only 0.25%, it was also significant ($t(4835) = 2.40$, $p < 0.01$, one-tailed). This is completely in line with our findings when using the full set of topics.

What happens when we take an agreement of a click among 3 different IPs? Again we focus on the MAP scores. We see that *BOOL* is again significantly the worst performing system ($p < 0.01$), and *OKAPI* is significantly performing better on a 1% significance level. Interestingly, we see that *LM* is slightly performing better than *LMS* ($p < 0.05$). As Table 4 shows that when we increase the agreement threshold, there are only 2,147 queries are left, which are predominantly very short (limiting the impact of smoothing) and many of them having having only a single relevant document. Finally, what happens when we take an agreement among 4 different IPs? We still see the same pattern as the previous runs, with *BOOL* being the worst, and *OKAPI* the best ($p < 0.01$). The findings are also consistent with the MRR and nDCG scores.

In summary, there are two implications. First, deriving a test collection using agreement of 2 is a viable alternative for using the whole log file. Second, the system rankings are consistent when treating the clicks as binary pseudo-relevance judgements (MAP, MRR) and as graded relevance judgements (nDCG).

5 External Validation

We investigate the validity of the log-based test collection in terms of the resulting system ranking. As ground-truth we use a test collection constructed by

Table 6. System-ranking of runs over e-mail topics

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1521 (5)	0.2632 (4)	0.3135 (3)	0.3147 (2)	0.3478 (1)
MRR	0.1732 (5)	0.3040 (4)	0.3550 (2)	0.3550 (2)	0.3907 (1)
nDCG	0.2430 (5)	0.3396 (4)	0.4386 (3)	0.4394 (2)	0.4623 (1)

```

<topic nr="34">
  <title>Frans Beelaerts Blokland Peking Beijing</title>
  <narrative>I am writing a book about foreigners in Beijing from the Boxer Rising
in 1900 to the Communist takeover 1949. Jonkheer Frans Beelaerts van Blokland was
the Dutch Minister in Peking during the World War One. I am very interested in seeing
any papers that you may hold relating to his years in Peking.</narrative>
  <files>2.05.90.xml ; 2.05.19.xml ; 2.21.253.xml</files>
</topic>

```

Fig. 1. An example of a topic based on an e-mail reference request

human experts: responses of archivists to e-mail reference questions. The system rankings of the log-based test collection are compared to this ground-truth.

5.1 Test Collection Based on E-Mail Reference Requests

We analyze a subset of e-mails that the NA-NL received from users, and with replies from archivists that referred explicitly to EAD files. We look at all correspondence (4.1GB of data). The e-mails are converted from PST file format to mbox format. Eventually, we manually derive 73 different topics (and recommended EAD links) from the e-mail files. A typical example is the information request in Fig. 1.

The explanation of the information request is included in `<narrative>`, the topic in `<title>`, and the relevant files for that topic in `<file>`. We selected typical replies from an archivist who linked to EAD files using the user query, or recommended the EAD finding aids which are relevant.

5.2 System Rank Correlations

We again use the Paired Samples T-test to check for significance by looking at the MAP scores. The results of Table 6 show that *BOOL* performs worst as well ($p < 0.01$, one-tailed). When we rank with LM without smoothing, there is also a significant improvement of 73.04% over *BOOL* ($t(67) = 3.22$, $p < 0.01$, one-tailed). When we use LM extended with smoothing, we see a 19.11% significant improvement. However, the difference between the LMS and NLLR models was only 0.38%, and is not significant. Moreover, *OKAPI* performed 10.52% better than *NLLR*, but is not significantly better. The findings are similar using the MRR and nDCG measures.

How reliable are our test collections derived from log files when compared to the test collection manually derived from e-mails and experts' replies? When we compare the system rankings of the test collection from the whole log (Table 3) with the e-mail topics (Table 6) using the MAP scores, we see a complete agreement with a Kendall's Tau value of 1. Overall, we see full agreement between the log-based evaluation and the reference requests, and high agreement among the test collections of the log-based evaluation approach.

6 Conclusions and Future Work

This paper investigated a search log-based approach to the evaluation of digital libraries. By using the DLs own collection and exploiting readily available interaction data in search logs, we can create a domain-specific test collection tailored to the case at hand. That is, having a representative document collection and representative sets of search requests. As a test collection, it can be used and reused for comparative testing under the same experimental conditions.

We conducted a large case study using a set of EAD documents and search logs of an archival institution. This resulted in a test collection to evaluate the retrieval of digital archives. This extends initial experiments using a museum's log file to create a domain-specific test collection [1], by using a massive archival collection from the National Archives of the Netherlands, and a massive search log covering several years of this high-profile website. We presented generic methods to derive a domain-specific test collection, and used a range of retrieval models to determine the effectiveness of the test collections. Our extraction methods are naive—we treat every clicked document as pseudo-relevant—but suffice to determine the viability of the approach. We validated the results against a set of traditional topics derived from email requests to the archive and the archivist's responses. We found complete agreement between the log-based evaluation and the traditional topics.

In our future work, we will further refine the log-based approach to evaluation, by using more advanced click models and by filtering out interesting categories of search requests. We will also use our test collections for improving archival access, by developing retrieval models for archival descriptions. These methods will be applicable to all archival institutions publishing their finding aids in EAD, the *de facto* standard. In addition, we are currently extracting the navigation within the archival finding aids, allowing us to locate particular document-components or parts of the archive of user interest. This allows us to build an evaluation set for focused or sub-document retrieval.

Acknowledgments. We gratefully thank Henny van Schie of the National Archives of the Netherlands for providing the data, and for valuable discussion and feedback on early versions of this paper. This research is supported by the Netherlands Organisation for Scientific Research (NWO) under project # 639.072.601.

References

- [1] Arampatzis, A., Kamps, J., Koolen, M., Nussbaum, N.: Deriving a domain specific test collection from a query log. In: *LaTeCH 2007*, pp. 73–80. ACL (2007)
- [2] Bailey, P., Craswell, N., Hawking, D.: Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.* 39(6), 853–871 (2003)
- [3] Boncz, P.A., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., Teubner, J.: *MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine*. In: *SIGMOD 2006*, pp. 479–490. ACM, New York (2006)
- [4] Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: *Extensible markup language (XML) 1.0*, 5th edn.(2008)
- [5] Dumais, S., Joachims, T., Bharat, K., Weigend, A.: SIGIR 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum* 37, 50–54 (2003)
- [6] Dupret, G., Liao, C.: A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In: *WSDM 2010*. ACM Press, New York (2010)
- [7] Feeney, K.: Retrieval of archival finding aids using world-wide-web search engines. *The American Archivist* 62(2), 206–228 (1999)
- [8] Fuhr, N., Tsakonias, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Sølvsberg, I.: Evaluation of digital libraries. *Int. J. on Digital Libraries* 8(1), 21–38 (2007)
- [9] Hiemstra, D., Rode, H., van Os, R., Flokstra, J.: PF/Tijah: text search in an XML database system. In: *OSIR 2006*, pp. 12–17 (2006)
- [10] Hutchinson, T.: Strategies for Searching Online Finding Aids: A Retrieval Experiment. *Archivaria* 44, 72–101 ((Fall 1997)
- [11] Jansen, B.J.: Search log analysis: What it is, what’s been done, how to do it. *Library & Information Science Research* 28(3), 407–432 (2006)
- [12] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: *SIGIR 2005*, pp. 154–161 (2005)
- [13] Jones, S., Cunningham, S.J., McNab, R.J., Boddie, S.J.: A transaction log analysis of a digital library. *Int. J. on Digital Libraries* 3(2), 152–169 (2000)
- [14] Lalmas, M.: *XML Information Retrieval*. *Encycl. of Lib. and Inf. Sciences* (2009)
- [15] Lytle, R.H.: Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval. *American Archivist* 43, 64–75 (Winter 1980)
- [16] Peters, T.: The history and development of transaction log analysis. *Library Hi Tech.* 42(11), 41–66 (1993)
- [17] Pitti, D.V.: Encoded Archival Description: An Introduction and Overview. *D-Lib Magazine* 5(11) (1999)
- [18] Ribeiro, F.: Subject Indexing and Authority Control in Archives: The Need for Subject Indexing in Archives and for an Indexing Policy Using Controlled Language. *Journal of the Society of Archivists* 17(1), 27–54 (1996)
- [19] Robertson, S.: On the history of evaluation in IR. *J. Inf. Sci.* 34(4), 439–456 (2008)
- [20] Shaw, W.M., Wood, J.B., Wood, R.E., Tibbo, H.R.: The cystic fibrosis database: content and research opportunities. *Library & Information Science Research* 13, 347–366 (1991)

- [21] Spärck Jones, K., van Rijsbergen, C.J.: Information retrieval test collections. *J. of Documentation* 32(1), 59–75 (1976)
- [22] Tibbo, H.R., Meho, L.I.: Finding finding aids on the world wide web. *The American Archivist* 64(1), 61–77 (2001)
- [23] White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: *SIGIR 2007*, pp. 255–262. ACM, New York (2007)
- [24] Zhang, J., Kamps, J.: Searching archival finding aids: Retrieval in original order? In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) *ECDL 2009. LNCS*, vol. 5714, pp. 447–450. Springer, Heidelberg (2009)
- [25] Zhang, J., Kamps, J.: Focused search in digital archives. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) *WISE 2009. LNCS*, vol. 5802, pp. 463–471. Springer, Heidelberg (2009)

Determining Time of Queries for Re-ranking Search Results*

Nattiya Kanhabua and Kjetil Nørøvåg

Dept. of Computer Science,
Norwegian University of Science and Technology,
Trondheim, Norway

Abstract. Recent work on analyzing query logs shows that a significant fraction of queries are temporal, i.e., relevancy is dependent on time, and temporal queries play an important role in many domains, e.g., digital libraries and document archives. Temporal queries can be divided into two types: 1) those with temporal criteria explicitly provided by users, and 2) those with no temporal criteria provided. In this paper, we deal with the latter type of queries, i.e., queries that comprise *only keywords*, and their relevant documents are associated to particular time periods not given by the queries. We propose a number of methods to determine the time of queries using temporal language models. After that, we show how to increase the retrieval effectiveness by using the determined time of queries to re-rank the search results. Through extensive experiments we show that our proposed approaches improve retrieval effectiveness.

1 Introduction

An enormous amount of information is stored in the form of digital documents, examples include web pages harvested and stored in web archives as well as newspaper articles stored in news archives. Information in such document repositories are useful for both expert users, e.g., historians, librarians, and journalists, as well as for students and other people searching for information needs. However, when searching in such *temporal document collections*, it is difficult to achieve high accuracy using simple keyword search because the contents are strongly time-dependent; documents are about events that happened at a particular time period. In addition, accesses to the contents are time-dependent too, i.e., time is a part of the information needs represented by temporal queries.

In previous work [210], searching temporal document collections has been performed by issuing temporal queries composed of keywords, and the creation or update date of documents (called temporal criteria). In that way, a system narrows down the results by retrieving documents according to both text and temporal criteria. Temporal queries can be divided into two categories: 1) those with temporal criteria explicitly provided by users [210], and 2) those with no temporal criteria provided. An example of a query with temporal criteria explicitly provided is “the U.S. Presidential election 2008”, while that of a query without temporal criteria provided is “Germany FIFA

* This work has been supported by the LongRec project, partially funded by the Norwegian Research Council.

World Cup”. However, for the latter example, a user’s temporal intent is *implicitly* provided, i.e., referring to the world cup event in 2006. As mentioned in [1], an analysis of web user query log shows that 1.5% of queries are explicitly provided with temporal criteria [11], i.e., containing temporal expressions, while about 7% of web queries have temporal intent implicitly provided [9].

In this paper, we focus on implicit temporal queries, i.e., temporal queries that comprise *only keywords*, and where relevant documents are associated to particular time periods that are not given by the queries. Through a novel approach for determining the time of queries, or implicit temporal intent, using temporal language models, we are able to increase the retrieval effectiveness by using the determined time of queries to re-rank search results. Thus, the main contributions of this paper are: 1) the first study on how to determine the time of queries without temporal criteria provided, as well as techniques for determining this time, 2) a study on how to incorporate the determined time of queries into the re-ranking search results, and 3) an extensive evaluation of our approaches for determining the time of queries, as well as of re-ranking search results using the time of queries. It should be noted that our approach is language-independent: the only requirement is the availability of a temporal document collection in the query language, such a corpus can be easily obtained from, for example, a news archive.

The organization of the rest of the paper is as follows. In Sect. 2 we give an overview of related work. In Sect. 3 we outline our of document and query models. Then, we explain the use of temporal language models for document dating. In Sect. 4 we present our approaches to determining the time of queries without temporal criteria provided. In Sect. 5 we describe how to use the determined time to improve the retrieval effectiveness. In Sect. 6 we evaluate our proposed query dating, and re-ranking methods. Finally, in Sect. 7 we conclude and outline our future work.

2 Related Work

Recently, a number of papers have described issues of temporal search [2,10,13]. In the approaches described in [2,10], a user explicitly specifies time as a part of query. Typically, such a temporal query is composed of query keywords and temporal criteria, which can be a point in time or a time interval. In general, temporal ranking can be divided into two types: approaches based on *link-based analysis* and *content-based analysis*. The first approach studies link structures of a document and uses this information in a ranking process, whereas the second approach examines the contents of a document instead of links. In our context, we will focus on analyzing contents only because information about links is not available in all domains, and content-based analysis seems to be more practical for a general search application. Temporal ranking exploiting document contents and temporal information are presented in [4,5,8,12,13].

In [8], Li and Croft incorporated time into language models, called time-based language models, by assigning a document prior using an exponential decay function of a document creation date. They focused on recency queries, such that the more recent documents obtain the higher probabilities of relevance. In [4], Diaz and Jones also used document creation dates to measure the distribution of retrieved documents and create the temporal profile of a query. They showed that the temporal profile together with the

contents of retrieved documents can improve average precision for the query by using a set of different features for discriminating between temporal profiles. In [13], Sato et al. defined a temporal query and proposed ranking taking into account time for fresh information retrieval. In [5] an approach to rank documents by freshness and relevance is presented. In [12], Perkiö et al. introduced a process of automatically detecting a topical trend (the strength of a topic over time) within a document corpus by analyzing the temporal behavior of documents using a statistic topic model.

Dating of documents has been previously studied by de Jong et al. [3], and their approach later extended by Kanhabua and Nørnvåg [6]. However, dating short queries and employing the time in ranking has to our knowledge not been performed before.

The most related work to this paper is [19]. Berberich et al. [1] integrated temporal expressions into query-likelihood language modeling, which considers uncertainty inherent to temporal expressions in a query and documents, i.e., temporal expressions can refer to the same time interval even they are not exactly equal. The work by Berberich et al. and our work is similar in the sense that both incorporate time into a ranking in order to improve the retrieval effectiveness for temporal search, however, in their work, the temporal criteria are explicitly provided for a query. Metzler et al. [9] also consider implicit temporal needs in queries. They proposed mining query logs and analyze query frequencies over time in order to identify strongly time-related queries. In addition, they presented a ranking concerning implicit temporal needs, and the experimental results showed that their approach improved the retrieval effectiveness of temporal queries for web search. Rather than relying on user query logs, we propose an alternative for determining the time of queries from the contents.

3 Preliminaries

In this section, we first briefly outline our document and query models. Then, we explain the basic approach to document dating using temporal language models.

3.1 Temporal Document Model

In this paper, a document collection contains a number of corpus documents defined as $C = \{d_1, \dots, d_n\}$. A document d_i can be seen as bag-of-words (an unordered list of terms), and a creation or updated date. Note that, d_i can also be associated to temporal expressions containing in the contents. However, temporal expressions will not be studied in this paper. Let $Time(d_i)$ be a function that gives a creation or updated date of d_i , so d_i can be represented as $d_i = \{\{w_1, \dots, w_n\}, Time(d_i)\}$. If C is partitioned wrt. a time granularity of interest, the associated time partition of d_i is a time interval $[t_k, t_{k+1}]$ containing $Time(d_i)$, that is $Time(d_i) \in [t_k, t_{k+1}]$. For example, if we partition C using the 1-month granularity and $Time(d_i)$ is 2010/03/05, the associated time partition of d_i will be [2010/03/01, 2010/03/31].

3.2 Temporal Query Model

We define a temporal query q as composed of two parts: keywords q_{word} and temporal criteria q_{time} , where $q_{word} = \{w_1, \dots, w_m\}$, and $q_{time} = \{t'_1, \dots, t'_l\}$ where t'_j is a time

interval, or $t'_j = [t_j, t_{j+1}]$. In other words, q contains uncertain temporal intent that can be represented by one or more time intervals. We can refer to q_{word} as topical features, and q_{time} as temporal features of q . Hence, our aim is to retrieve documents about the topic of query where their creation dates are corresponding to time criteria.

Recall that temporal queries can be divided into two types: 1) those with temporal criteria explicitly provided by a user, and 2) those with no temporal criteria provided. An example of the first type is “Summer Olympics 2008” where the user interests in documents about “Summer Olympics” written in 2008. In this case, q_{time} is equal to $\{[2008/01/01, 2008/12/31]\}$ given the 1-year time granularity. Queries in the second type can be implicitly associated with particular time especially queries related to periodic, or outbreak events. The query “Boxing Day tsunami” is associated with the year “2004”, $q_{time} = \{[2004/01/01, 2004/12/31]\}$, and the query “the U.S. presidential election” can be associated with the years “2000”, “2004”, and “2008”, so that $q_{time} = \{[2000/01/01, 2000/12/31], \dots, [2008/01/01, 2008/12/31]\}$. When the time q_{time} is not given explicitly by the user, it has to be determined by the system, as will be described later in this paper.

3.3 Temporal Language Models

The document dating approach is based on the *temporal language model* presented in [3], which is a variant of the time-based model in [8]. The idea is to assign a probability to a time partition according to word usage or word statistics over time.

A normalized log-likelihood ratio [7] is used to compute the similarity between two language models. Given a partitioned corpus, it is possible to determine the timestamp of a non-timestamped document d_i by comparing the language model of d_i with each corpus time partition p_j using the following equation:

$$Score(d_i, p_j) = \sum_{w \in d_i} P(w|d_i) \times \log \frac{P(w|p_j)}{P(w|C)} \quad (1)$$

where C is the background model estimated on the entire collection. Smoothing will be employed to avoid the zero probability of unseen words. The timestamp of the document is the time partition maximizing a score according to the equation above.

In order to build temporal language models, a temporal corpus is needed. The temporal corpus can be any document collection where 1) the documents are timestamped with creation time, 2) covering a certain time period (at least the period of the queries collections), and 3) containing enough documents to make robust models. A good basis for such a corpus is a news archive. We will use the *New York Times annotated corpus* [1] since it is readily available for research purposes. However, any corpus with similar characteristics can be employed, including non-English corpora for performing dating of non-English texts. We will in the following denote a temporal corpus as \mathcal{D}_N .

¹ http://www.ldc.upenn.edu/Catalog/docs/LDC2008T19/new_york_times_annotated_corpus.pdf

4 Determining Time of Queries Using Temporal Language Models

In this section, we describe three approaches to determining the time of queries when no temporal criteria are provided. The first two approaches use temporal language models (cf. Sect. 3) as basis, and the last approach uses no language models. The first approach performs dating queries using keywords only. The second approach takes into account the fact that in general queries are short, and aims at solving this problem with a technique inspired by pseudo-relevance feedback (PRF) that uses the *top-k* retrieved documents in dating queries. The third approach also uses the *top-k* retrieved documents by PRF and assumes their creation dates as the time of queries.

All approaches will return a set of determined time intervals and their weights, which will be used in re-ranking documents in order to improve the retrieval effectiveness as described in more detail in Sect. 5.

4.1 Dating Query Using Keywords

Our basic technique for query dating is based on using keywords only, and it is described formally in Algorithm 1.

The first step is to build temporal language models T_{LM} from the temporal document corpus (line 5), which essentially is the statistics of word usage (raw frequencies) in all time intervals, which are partitioned wrt. the selected time granularity g . Table 1 illustrates a subset of temporal language models. Creating the temporal language models (basically aggregating statistics grouped on time periods) is obviously a costly process, and will be done just once as an off-line process and then only the statistics have to be retrieved at query time.

For each time partition p_j in T_{LM} , the similarity score between q_{word} and p_j is computed (line 7). The similarity score is calculated using a normalized log-likelihood ratio according to Equation 1. Each time partition p_j and its computed score will be stored in C , or the set of time intervals and scores (line 8). After computing the scores for all time partitions, the contents of C will be sorted by similarity score, and then the *top-m* time intervals are selected as the output set A (line 10).

Finally, the determined time intervals resulting from Algorithm 1 will be assigned weights indicating their importance. In our approach, we simply give a weight to each time interval using its reverse ranked number. For example, if the output set A contains top-5 ranked time intervals, the intervals ranked 1, 2, 3, 4, and 5 will have the weights 5, 4, 3, 2, and 1 respectively.

Table 1. Example of contents of temporal language models

Time	Term	Frequency
2001	World Trade Center	1545
2002	Terrorism	2236
2003	Iraq	1510
2004	Euro 2004	750
2004	Athens	1213
2005	Terrorism	1990
2005	Tsunami	3528
2005	Hurricane Katrina	1012
2008	Obama	2030

Algorithm 1. *DateQueryKeywords*($q_{\text{word}}, g, m, \mathcal{D}_{\mathcal{N}}$)

```

1: INPUT: Query  $q_{\text{word}}$ , time granularity  $g$ , number of time intervals  $m$ , and temporal corpus  $\mathcal{D}_{\mathcal{N}}$ 
2: OUTPUT: Set of time intervals associated to  $q_{\text{word}}$ 
3:  $A \leftarrow \emptyset$  // Set of time intervals
4:  $C \leftarrow \emptyset$  // Set of time intervals and scores
5:  $T_{LM} \leftarrow \text{BuildTemporalLM}(g, \mathcal{D}_{\mathcal{N}})$ 
6: for each  $\{p_j \in T_{LM}\}$  do
7:    $\text{score}_{p_j} \leftarrow \text{CalSimScore}(q_{\text{word}}, p_j)$  // Compute similarity score of  $q_{\text{word}}$  and  $p_j$ 
8:    $C \leftarrow C \cup \{(p_j, \text{score}_{p_j})\}$  // Store  $p_j$  and its similarity score
9: end for
10:  $A \leftarrow C.\text{selectTopMIntervals}(m)$  // Select top- $m$  intervals ranked by scores
11: return  $A$ 

```

4.2 Dating a Query Using Top-k Documents

In our second approach to query dating, the idea is that instead of dating query keywords q_{word} directly, we will instead date the *top-k* retrieved documents of the (non-temporal) query q_{word} . The resulting time of the query will be the combination of determined times of each top-k document.

The algorithm for dating a query using top-k retrieved documents is given in Algorithm 2. First, we retrieve documents by issuing a (non-temporal) query q_{word} , and retrieve only the *top-k* result documents (line 5). Then, temporal language models T_{LM} are built as described previously (line 6). For each document d_i in $D_{\text{Top}K}$, compute its similarity score with each time partition p_j in T_{LM} (lines 10-13). After computing scores for d_i for all time partitions, sort the contents of C by similarity score, and select only *top-m* time intervals as the results of d_i (line 14).

The next step is to update the set B with a set of time results C_{tmp} obtained from dating d_i . This is performed as follows: For each time interval p_k in C_{tmp} , check if B already contains p_k (line 16). If p_k exists in B , get a frequency of p_k and increase the frequency by 1 (lines 17-18). If p_k does not exist in B , add p_k into B as a new time interval and set its frequency to 1 (line 20). After dating all documents in $D_{\text{Top}K}$, sort the contents of B by frequency, and select only the *top-m* time intervals as the output set A (line 25).

The weights of time intervals will be their reverse ranked number. Note that it can be only one time interval in each rank of an output obtained from Algorithm 1, while it can be more than one time interval in each rank in case of Algorithm 2.

4.3 Using Timestamp of Top-k Documents

The last approach is a variant of the dating using *top-k* documents described above. The idea is similar in the use of the *top-k* retrieved documents of the (non-temporal) query q_{word} . The resulting time of the query will be the creation date (or timestamps) of each top-k document. In this case, no temporal language models are used.

Algorithm 2. *DateQueryWithTopkDoc*($q_{\text{word}}, g, m, k, \mathcal{D}_{\mathcal{N}}$)

```

1: INPUT: Query  $q_{\text{word}}$ , time granularity  $g$ , number of intervals and documents  $m, k$ , temporal corpus  $\mathcal{D}_{\mathcal{N}}$ 
2: OUTPUT: Set of time intervals associated to  $q_{\text{word}}$ 
3:  $A \leftarrow \emptyset$  // Set of time intervals
4:  $B \leftarrow \emptyset$  // Set of time intervals and their frequencies
5:  $D_{\text{TopK}} \leftarrow \text{RetrieveTopKDoc}(q_{\text{word}}, k)$  // Retrieve top-k documents
6:  $T_{\text{LM}} \leftarrow \text{BuildTemporalLM}(g, \mathcal{D}_{\mathcal{N}})$ 
7: for each  $\{d_i \in D_{\text{TopK}}\}$  do
8:    $C \leftarrow \emptyset$  // Set of time intervals and scores
9:    $C_{\text{imp}} \leftarrow \emptyset$  // Set of time intervals
10:  for each  $\{p_j \in T_{\text{LM}}\}$  do
11:     $\text{score}_{p_j} \leftarrow \text{CalSimScore}(d_i, p_j)$  // Compute similarity score of  $d_i$  and  $p_j$ 
12:     $C \leftarrow C \cup \{(p_j, \text{score}_{p_j})\}$  // Store  $p_j$  and its similarity score
13:  end for
14:   $C_{\text{imp}} \leftarrow C.\text{selectTopMIntervals}(m)$  // Select top-m intervals by scores
15:  for each  $\{p_k \in C_{\text{imp}}\}$  do
16:    if  $B$  has  $p_k$  then
17:       $\text{freq} \leftarrow B.\text{getFreqForTInterval}(p_k)$  // Get frequency of  $p_k$ 
18:       $B \leftarrow B.\text{updateFreqForTInterval}(p_k, \text{freq} + 1)$  // Increase frequency by 1
19:    else
20:       $B \leftarrow B.\text{addTInterval}(p_k, 1)$  // Add a new time interval and set its frequency to 1
21:    end if
22:  end for
23: end for
24:  $A \leftarrow B.\text{selectTopMIntervals}(m)$  // Select top-m intervals ranked by frequency
25: return  $A$ 

```

5 Re-ranking Documents Using the Determined Time of Queries

In this section, we will describe how to use the time of queries determined by our approaches to improve the retrieval effectiveness. The idea is that, in addition to the documents' scores wrt. keywords, we will also take into account the documents' scores wrt. the *implicit* time of queries. Intuitively, documents with creation dates that closely match with the time of queries are more relevant and should be ranked higher.

There are a number of methods to combine a time score with existing text-based weighting models. For example, a time score can be combined with TF-IDF weighting using a linear combination, or it can be integrated into language modeling using a document prior probability as in [8]. In this paper, we propose to use a mixture model of a keyword score and a time score. Given a temporal query q with the determined time q_{time} , the score of a document d can be computed as follows:

$$S(q, d) = (1 - \alpha) \cdot S'(q_{\text{word}}, d_{\text{word}}) + \alpha \cdot S''(q_{\text{time}}, d_{\text{time}}) \quad (2)$$

where α is a parameter underlining the importance of a keyword score $S'(q_{\text{word}}, d_{\text{word}})$ and a time score $S''(q_{\text{time}}, d_{\text{time}})$. A keyword score $S'(q_{\text{word}}, d_{\text{word}})$ can be implemented using any of existing text-based weighting models, and it can be normalized as

$S'_{norm}(q_{word}, d_{word}) = \frac{S'(q_{word}, d_{word})}{\max S'(q_{word}, d_{word}, i)}$ where $\max S'(q_{word}, d_{word}, i)$ is the maximum keyword score among all documents.

For a time score $S''(q_{time}, d_{time})$, we formulate the probability of generating the time of query q_{time} given the associated time partition of document d_{time} as:

$$\begin{aligned} S''(q_{time}, d_{time}) &= P(q_{time} | d_{time}) \\ &= P(\{t'_1, \dots, t'_n\} | d_{time}) \\ &= \frac{1}{|q_{time}|} \sum_{t'_j \in q_{time}} P(t'_j | d_{time}) \end{aligned} \quad (3)$$

where q_{time} is a set of time intervals $\{t'_1, \dots, t'_n\}$ and $(t'_1 \cap t'_2 \cap \dots \cap t'_n) = \emptyset$. So, $P(q_{time} | d_{time})$ is an average of the probability of generating a time interval, or $P(t'_j | d_{time})$, over all the number of time intervals in q_{time} , or $|q_{time}|$.

The probability of generating a time interval t'_j given the time partition of document d_{time} can be defined in two ways as proposed in [11]: 1) ignoring uncertainty, and 2) taking uncertainty into account. By ignoring uncertainty, $P(t'_j | d_{time})$ is defined as:

$$P(t'_j | d_{time}) = \begin{cases} 0 & \text{if } d_{time} \neq t'_j, \\ 1 & \text{if } d_{time} = t'_j. \end{cases} \quad (4)$$

In this case, the probability of generating query time will be equal to 1 only if d_{time} is exactly the same as t'_j . By taking into account a weight of each time interval t'_j , $P(t'_j | d_{time})$ with *uncertainty-ignorant* becomes

$$P(t'_j | d_{time}) = \begin{cases} 0 & \text{if } d_{time} \neq t'_j, \\ \frac{w(t'_j)}{\sum_{t'_k \in q_{time}} w(t'_k)} & \text{if } d_{time} = t'_j. \end{cases} \quad (5)$$

where $w(t'_j)$ is a function giving a weight for a time interval t'_j , which is normalized by the sum of all weights $\sum_{t'_k \in q_{time}} w(t'_k)$.

In the case where uncertainty is concerned, $P(t'_j | d_{time})$ is defined using an exponential decay function:

$$P(t'_j | d_{time}) = DecayRate^{\lambda \cdot |t'_j - d_{time}|} \quad (6)$$

where $DecayRate$ and λ are constant, $0 < DecayRate < 1$ and $\lambda > 0$. Intuitively, this function gives a probability that decreases proportional to the difference between a time interval t'_j and the time partition of document d_{time} . A document with its time partition closer to t'_j will receive a higher probability than a document with its time partition farther from t'_j . By incorporating a weight of each time interval t'_j , $P(t'_j | d_{time})$ with *uncertainty-aware* becomes

$$P(t'_j | d_{time}) = \frac{w(t'_j)}{\sum_{t'_k \in q_{time}} w(t'_k)} \times DecayRate^{\lambda \cdot |t'_j - d_{time}|} \quad (7)$$

The normalization of $S''_{norm}(q_{time}, d_{time})$ can be computed in two ways: 1) uncertainty-ignorant using $P(t'_j | d_{time})$ defined in Equation 5 and 2) uncertainty-aware using

$P(t'_j | d_{time})$ defined in Equation 7. Finally, the normalized value of $S'_{norm}(q_{time}, d_{time})$ will be substituted $S''(q_{time}, d_{time})$ in Equation 8 yielding the normalized score of a document d given a temporal query q with determined time q_{time} as follows:

$$S_{norm}(q, d) = (1 - \alpha) \cdot S'_{norm}(q_{word}, d_{word}) + \alpha \cdot S''_{norm}(q_{time}, d_{time}) \quad (8)$$

6 Experiments

In this section, we will perform two experiments in order to evaluate our proposed approaches: 1) determining the time of queries using temporal language models, and 2) re-ranking search results using the determined time. In this section, we will describe the setting for each of the experiments, and then the results.

6.1 Experimental Setting

As we mentioned earlier, we can use any news archive collection to create temporal language models. In this paper, we used the New York Times annotated corpus as the temporal corpus. This collection contains over 1.8 million articles covering a period of January 1987 to June 2007. The temporal language models were created and stored in databases using Oracle Berkeley DB version 4.7.25.

To evaluate the query dating approaches, we obtained queries from Robust2004, which is a standard test collection for the TREC Robust Track containing 250 topics (topics 301-450 and topics 601-700). As reported in [8], some TREC queries favor documents in particular time periods. Similarly, we analyzed a distribution of relevant documents of the Robust2004 queries over time, and we randomly selected 30 strongly time-related queries (with the topic number: 302, 306, 315, 321, 324, 330, 335, 337, 340, 352, 355, 357, 404, 415, 428, 435, 439, 446, 450, 628, 648, 649, 652, 653, 656, 667, 670, 676, 683, 695). Time intervals of relevant documents were assumed as the correct time of queries. We measured the performance using precision, recall and F-score. Precision is the fraction of determined time intervals that are correct, while recall indicates the fraction of correct time intervals that are determined. F-score is the weighted harmonic mean of precision and recall, where we set $\beta = 2$ in order to emphasize recall. For query dating parameters, we used the top- m interval with $m = 5$, and the time granularity g and the top- k documents were variable in the experiments.

To evaluate the re-ranking approaches, the Terrier search engine was employed, and we used the BM25 probabilistic model with Generic Divergence From Randomness (DFR) weighting as our retrieval model. For the simplicity, we used default parameter settings for the weighting function. Terrier provides a mechanism to alter scores for retrieved documents by giving prior scores to the documents. In this way, we re-ranked search results at the end of retrieval by combining a keyword score $S'(q_{word}, d_{word})$ and a time score $S''(q_{time}, d_{time})$ as defined in Equation 8. We conducted re-ranking experiments using two collections: 1) the Robust2004 collection, and 2) the New York Times annotated corpus. For the Robust2004 collection, we used the 30 queries as temporal queries without time explicitly provided. The retrieval effectiveness of temporal search using the Robust2004 collection is measured by Mean Average Precision (MAP), and R-precision. For the New York Times annotated corpus, we selected 24 queries from a

Table 2. Example of the Google zeitgeist queries and associated time intervals

Query	Time	Query	Time
diana car crash	1997	madrid bombing	2005
world trade center	2001	pope john paul ii	2005
osama bin laden	2001	tsunami	2005
london congestion charges	2003	germany soccer world cup	2006
john kerry	2004	torino games	2006
tsa guidelines liquids	2004	subprime crisis	2007
athens olympics games	2004	obama presidential campaign	2008

Table 3. Query dating performance using precision, recall and F-score

Method	Precision		Recall		F-score($\beta = 2$)	
	6-month	12-month	6-month	12-month	6-month	12-month
QW	.56	.67	.34	.64	.37	.65
PRF ($k=5$)	.55	.63	.47	.79	.48	.75
PRF ($k=10$)	.56	.60	.46	.74	.48	.71
PRF ($k=15$)	.54	.60	.42	.70	.44	.68
NLM ($k=5$)	.92	.97	.35	.44	.40	.49
NLM ($k=10$)	.90	.95	.48	.56	.53	.61
NLM ($k=15$)	.89	.93	.56	.63	.61	.67

historical collection of aggregated search queries, or the Google zeitgeist². An example of temporal queries are shown in Table 2. The temporal searches were conducted by human judgment. Performance measures are the precision at 5, 10, and 15 documents, or P@5, P@10, and P@15 respectively. For re-ranking parameters, we used an exponential decay rate $DecayRate = 0.5$, and $\lambda = 0.5$. A mixture model parameter was obtained from the experiments, where $\alpha = 0.05$ and 0.10 for *uncertainty-ignorant* and *uncertainty-aware* methods respectively.

The description of different approaches is given as follows. **QW** determines time using keywords *plus* uncertainty-ignorant re-ranking. **QW-U** determines time using keywords *plus* uncertainty-aware re-ranking. **PRF** determines time using top-k retrieved documents *plus* uncertainty-ignorant re-ranking. **PRF-U** determines time using top-k retrieved documents *plus* uncertainty-aware re-ranking. **NLM** assumes creation dates of top-k retrieved documents as the time of queries (no language models used) *plus* uncertainty-ignorant re-ranking. **NLM-U** assumes creation dates of top-k retrieved documents as the time of queries (no language models used) *plus* uncertainty-aware re-ranking. Top-k documents were retrieved using pseudo relevance feedback, i.e., the result documents after performing query expansion using Rocchio algorithm.

6.2 Experimental Results

The performance of query dating methods are shown in Table 3. NLM performs best in precision for all time granularities whereas PRF performs best in recall (only for 12-month). NLM and PRF give the best F-score results for 6-month and 12-month respectively. In general, the smaller k tends to give the better results, while 12-month

² <http://www.google.com/intl/en/press/zeitgeist/index.html>

Table 4. Re-ranking performance using MAP and R-precision with the baseline performance 0.3568 and 0.3909 respectively (the Robust2004 collection)

Method	MAP		R-precision	
	6-month	12-month	6-month	12-month
QW	.3565	.3576	.3897	.3924
QW-U	.3556	.3573	.3925	.3943
PRF ($k=5$)	.3564	.3570	.3885	.3926
PRF ($k=10$)	.3568	.3570	.3913	.3919
PRF ($k=15$)	.3566	.3567	.3912	.3921
PRF-U ($k=5$)	.3548	.3574	.3903	.3950
PRF-U ($k=10$)	.3538	.3576	.3904	.3935
PRF-U ($k=15$)	.3538	.3572	.3893	.3940
NLM ($k=5$)	.3585	.3589	.3924	.3917
NLM ($k=10$)	.3586	.3591	.3918	.3925
NLM ($k=15$)	.3584	.3596	.3898	.3934
NLM-U ($k=5$)	.3604	.3608	.3975	.3978
NLM-U ($k=10$)	.3604	.3610	.3953	.3961
NLM-U ($k=15$)	.3606	.3620	.3943	.3967

Table 5. Re-ranking performance using P@5, P@10, and P@15 with the baseline performance 0.35, 0.30 and 0.27 respectively * indicates statistically improvement over the baselines using t-test with significant at $p < 0.05$ (the NYT collection)

Method	P@5		P@10		P@15	
	6-month	12-month	6-month	12-month	6-month	12-month
QW	.42	.45	.37	.39	.32	.33
QW-U	.40	.42	.35	.36	.30	.32
PRF ($k=15$)	.42	.46	.38	.42	.35	.39
PRF-U ($k=15$)	.41	.45	.36	.40	.33	.37
NLM ($k=15$)	.50	.52	.47	.49	.42	.44
NLM-U ($k=15$)	.53	.55*	.48	.50*	.45	.46*

yields higher performance compared to *6-month*. Finally, the performance of QW seems to be robust for *12-month* regardless of dating solely short keywords.

To evaluate re-ranking, the baseline of our experiments is a retrieval model without taking into account the time of queries, i.e., pseudo relevance feedback using Rocchio algorithm. For the Robust2004 queries, the baseline performance are MAP=0.3568 and R-precision=0.3909. Experimental results of MAP and R-precision are shown in Table 4. The results show that QW, QW-U, PRF, PRF-U outperformed the baseline in both MAP and R-precision for *12-month*, and NLM, NLM-U outperformed the baseline in all cases. PRF-U always performed better than PRF in both MAP and R-precision for *12-month*, while QW-U performed better than QW in R-precision for *12-month* only. NLM, NLM-U always outperformed the baseline and the other proposed approaches because using the creation dates of documents is more accurate than those obtained from the dating process. This depicts that taking time into re-ranking can better the

retrieval effectiveness. Hence, if query dating is improved with a high accuracy, the retrieval effectiveness will be improved significantly.

The results of evaluate the Google zeitgeist queries are shows in Table 5. In this case, we fix the number of *top-k* to 15 only. Table 5 illustrated the precision at 5, 10 and 15 documents. The baseline performance is $P@5=0.35$, $P@10=0.30$ and $P@15=0.27$. The results show that our proposed approaches perform better than the baseline in all cases. NLM, NLM-U performs the best among all proposed approaches.

7 Conclusions and Future Work

In this paper, we have studied implicit temporal queries where no temporal criteria is provided, and how to increase retrieval effectiveness for such queries. The effectiveness has been improved by determining the implicit time of the queries and employing this to re-rank the query results. Through extensive experiments we show that our proposed approach improves retrieval effectiveness.

Although using our approach shows improvement on retrieval effectiveness, the quality of the actual query dating processing is a limitation when aiming at further increase in effectiveness. Future work includes further improvement on the query dating based on external knowledge from sources like Wikipedia.

References

1. Berberich, K., Bedathur, S., Alonso, O., Weikum, G.: A language modeling approach for temporal information needs. In: Proceedings of ECIR 2010 (2010)
2. Berberich, K., Bedathur, S.J., Neumann, T., Weikum, G.: A time machine for text search. In: Proceedings of SIGIR 2007 (2007)
3. de Jong, F., Rode, H., Hiemstra, D.: Temporal language models for the disclosure of historical text. In: Proceedings of AHC 2005 (History and Computing) (2005)
4. Diaz, F., Jones, R.: Using temporal profiles of queries for precision prediction. In: Proceedings of the 27th SIGIR (2004)
5. Jatowt, A., Kawai, Y., Tanaka, K.: Temporal ranking of search engine results. In: Ngu, A.H.H., Kitsuregawa, M., Neuhold, E.J., Chung, J.-Y., Sheng, Q.Z. (eds.) WISE 2005. LNCS, vol. 3806, pp. 43–52. Springer, Heidelberg (2005)
6. Kanhabua, N., Nørvåg, K.: Improving temporal language models for determining time of non-timestamped documents. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 358–370. Springer, Heidelberg (2008)
7. Kraaij, W.: Variations on language modeling for information retrieval. SIGIR Forum 39(1), 61 (2005)
8. Li, X., Croft, W.B.: Time-based language models. In: Proceedings of CIKM (2003)
9. Metzler, D., Jones, R., Peng, F., Zhang, R.: Improving search relevance for implicitly temporal queries. In: Proceedings of SIGIR 2009 (2009)
10. Nørvåg, K.: Supporting temporal text-containment queries in temporal document databases. Journal of Data & Knowledge Engineering 49(1), 105–125 (2004)
11. Nunes, S., Ribeiro, C., David, G.: Use of temporal expressions in web search. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 580–584. Springer, Heidelberg (2008)
12. Perkiö, J., Buntine, W., Tirri, H.: A temporally adaptive content-based relevance ranking algorithm. In: Proceedings of the 28th SIGIR (2005)
13. Sato, N., Uehara, M., Sakai, Y.: Temporal ranking for fresh information retrieval. In: Proceedings of the 6th IRAL (2003)

Ranking Entities Using Web Search Query Logs

Bodo Billerbeck¹, Gianluca Demartini²,
Claudiu S. Firan², Tereza Iofciu², and Ralf Krestel²

¹ Microsoft Research, 7 JJ Thomson Avenue, Cambridge CB3 0FB, UK
bodob@microsoft.com

² L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany
{iofcui,demartini,firan,krestel}@L3S.de

Abstract. Searching for entities is an emerging task in Information Retrieval for which the goal is finding well defined entities instead of documents matching the query terms. In this paper we propose a novel approach to Entity Retrieval by using Web search engine query logs. We use Markov random walks on (1) Click Graphs – built from clickthrough data – and on (2) Session Graphs – built from user session information. We thus provide semantic bridges between different query terms, and therefore indicate meaningful connections between Entity Retrieval queries and related entities.

1 Introduction

Current Web search engines retrieve textually relevant Web pages for a given keyword query. The idea behind *Entity Retrieval* (ER) is to find entities directly. As an example, consider the ER query “hybrid cars” where relevant results would be *Toyota Prius* or *Honda Insight*, but not an informative page about hybrid vehicles. Instead of the user browsing through all Web pages retrieved by the search engine, a list of relevant entities should be presented to the user. As shown in previous work, a big percentage of web search engine queries are about entities [1]. A commercial product addressing such type of queries is Google Squared [2] where the results for queries such as “hybrid cars” is a table with instances of the desired type.

By mining a very large Web search engine query log with clickthrough data and session information we are able to create two types of graphs on which we can afterwards apply our algorithms: (1) We create a *Click Graph* by using queries and URLs as nodes and connecting and weighting them by their user click frequencies, and (2) A *Session Graph* by using only queries as nodes with edges between them if they appear in the same user sessions, again weighted by co-occurrence frequencies. In order to utilize this information source for improving ER we perform a Markov random walk on the graphs. We employ graph traversal techniques with different weighting schemes in order to match result entities to given queries. Experimental results show that the intersection of the click graph and the session graph is the best evidence for answering ER queries when traversing the graphs.

¹ <http://www.google.com/squared>

2 Related Work

Finding entities on the Web is a recent topic in the field of Information Retrieval. The first proposed approaches [3,4] mainly focus on scaling efficiently on Web dimension datasets but not on the effectiveness of search. In more detail, the authors of [4] tackle the ER task with a two component approach: one for extracting entities from the web and one for querying the database containing the extracted entities. A semantic search engine based on SPARQL queries, an optimized index structure, and an ontology is described in [3]. The system is implemented using YAGO([13]), a Wikipedia and WordNet based ontology, and Wikipedia itself as a corpus. The main differences of the above mentioned systems to our approach are that the user has to follow certain rules for querying the system; either stating the entity type that they are looking for or even some more complex structure requirements to transform the query into a SPARQL representation. We do not make any assumptions about the user query facilitating the interaction considerably. We also do not limit our system to certain entity types and use the Web as a corpus instead of e.g. Wikipedia.

In the wake of the INEX[2,7] challenge a couple of systems were presented to solve Entity Ranking in the Wikipedia context. Different strategies were used by the participants: The authors of [12] use link information on the Wikipedia pages; [9] make use of the category information present in Wikipedia and incorporate an ontology to improve effectiveness; [8] use NLP techniques; [14] leverages user provided example entities. A probabilistic framework for ER is proposed in [2].

Our ER algorithms exploit graph structures. Session Graphs or Click Graphs were previously used beneficially in various tasks. In [11] the authors perform an analysis of web search query logs and user activities concluding that 50% of queries are about entities. A probabilistic approach for named entity recognition in queries is presented in [10]. In [5] the authors describe how to use a Click Graph to improve Web search. In [6] session data is used to generate query suggestions. User session information is also used in [1] for improving Web search results. In our work we apply a Markov random walk model on both Click and Session Graphs and investigate its use for answering Entity Search tasks.

3 Constructing and Entering the Graphs

The Click Graph. A click log consists of a set of URLs $U = u_1, \dots, u_n$ that users clicked on in response to queries $Q = q_1, \dots, q_n$. Our approach for constructing the graphs is based on previous work of Craswell and Szummer [5]. We can build a *click graph* based on the notion of co-clicked URLs. In a click graph each unique query (i.e., a string of keywords) q_i and each URL u_j is a node. We define the set of nodes $V \equiv Q \cup U$. There is a directed edge between a query node q_i and a URL u_j if at least one user clicked u_j in the result page of the query q_i . Moreover, there is a weight on each edge computed based on the number of times u_j was clicked as result of query q_i . Such a graph represents relations between

² <http://www.inex.otago.ac.nz/>

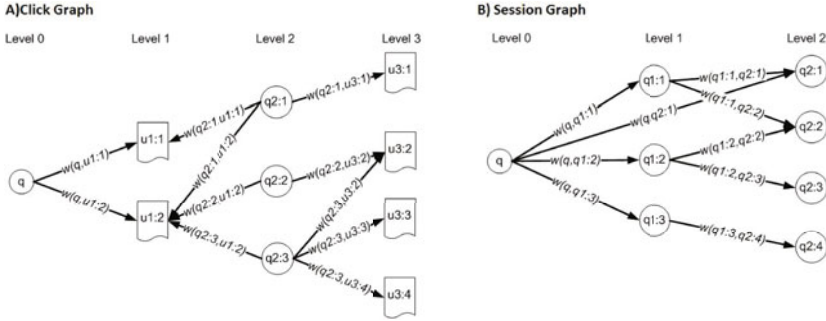


Fig. 1. Schemes of a Click Graph (A), connecting an ER query q with entities $q_{l,i}$ via URLs $U_{l,j}$ where l indicates the level, and of a Session Graph (B) connecting a ER query q with queries $q_{2,i}$ on level 2

queries and web documents as well as between different queries. We define q as the starting point for such search for entities: this is the ER query provided by the user (more details on how to properly select q are given in Section 3). We then assume queries close to q in the graph to be possible answers, that is, relevant entities q_i . In this way we can follow edges starting from q looking for relevant results (see Figure 1A).

The Session Graph. In a session graph nodes are formed by the set of queries $V \equiv Q = q_1, \dots, q_n$. There is a directed link from a query q_i to a query q_j if the query q_j was issued after query q_i in the same search session. Similarly, we can define q as the starting point, that is, the user’s ER query. We can then follow the edges looking for relevant results (that is, queries q_i) in the queries connected to q (see Figure 1B). Finally, the task of finding entities can be then defined as ranking queries q_i by probability of being relevant to the ER query q . The hypothesis is that a user posing an ER query which does not yield satisfying results will reformulate the query to find useful information. Upon inspection, it seems that the reformulated query often consists of an instance of the group of entities the user is looking for, e.g. “Spanish fish dishes” and “Paella”.

Finding the Entry Point in the Graph. We investigate how we can identify a suitable subset of logged queries from which entities related to a particular topic can be extracted. We describe a possible way of selecting q (i.e., the starting point of the random walk) given the ER query issued by the user. We search the user query in the available query log and use such query as the node q . For instance, the query “salad recipes” can be found in the click graph as depicted in Figure 2. We then perform a random walk from this node in the graph. Beginning from this query, at the distance of two nodes out, the random walk finds such queries as “chicken salad recipe” as well as “pasta salad”. Further out, the queries “green pea salad” and “caesar salad” are encountered. Specifically, we show the top ten queries with the highest transition probabilities from the node of origin (excluding the starting point), and a further five queries connected to two of these. While most of the queries directly linked to the original query are potentially useful for extracting entities, there are some queries that are less

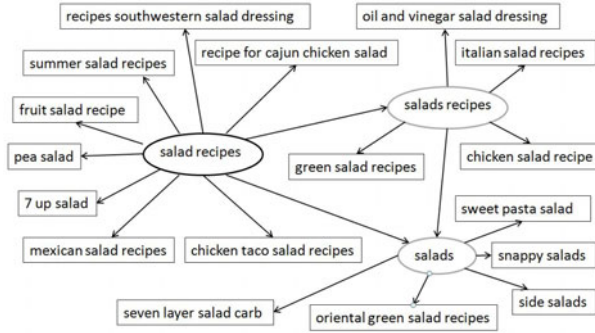


Fig. 2. Selection of walked queries for the query “salad recipes”

suitable for this task. However, these can be understood as categorising queries that may lead to other promising queries which may otherwise not be reached from the originating node. Examples of these ‘bridging queries’ are the nodes “salads” and “salads recipes” – singled out in Figure 2.

4 Walking the Graphs for Entity Ranking

Similarly to [5] we perform a Markov random walk on the click and session graphs in order to find relevant results for query q . The main difference is that our goal is to rank queries connected to q rather than ranking URLs by the probability distribution computed with the random walk. Moreover, the resulting entities are found only in the log queries, disregarding the text of the Web pages pointed to by the URLs in the log.

We define transition probabilities from a node j to a node k based on the click counts (i.e., $w(j, k)$ in Figure 1A and B) as $P_{t+1|t}(k|j) = \frac{w(j,k)}{\sum_i w(j,i)}$ where i ranges over all nodes connected to j . The notation $P_{t+1|t}(k|j)$ defines the probability of moving from node j at step t to node k at step $t + 1$.

By storing these single step transition probabilities in a matrix A where $A[j, k] = P_{t+1|t}(k|j)$, it is possible to compute a random walk of t steps starting from node j ($P_{t|0}(k|j)$) as $[A^t]_{jk}$. That is, we sum weights on the edges encountered on all paths of length t between the node j and a node k . The more paths the higher the random walk probability of reaching k starting from j .

4.1 Approaches Used on the Click Graph and Session Graph

At search time, the given ER query is matched in the graph and set as starting node (see Section 3). Performing a random walk over the graph, using query-URL-query transitions associated with weights on the edges (i.e. click frequencies), as shown in Figure 1A, enables us to find relevant entities as other queries in the graph and present them as a ranked list of entity results. We retrieve all queries reached within up to ten random walk steps in the click graph (i.e. five queries deep) and five steps in the session graph from the original query. The

retrieved set of results is ranked and/or filtered by one of the following methods and only results appearing two steps away (i.e. one query deep) from the original query are kept as precision values drop rapidly when considering more levels.

Simple Random Walk. This approach ranks all reached queries (interpreted as potential entities) by their random walk probability computed as described in Section 4 (using 0 as self transition probability and only forward walks) but keeps only queries which are one URL away from the original query (i.e., level 2 in Figure 1A) for the method labelled C_2 . For the method labelled C_{10} , we keep any queries encountered up to 10 steps away from the original queries. The result queries (potential entities) are ranked by their random walk transition scores over all possible paths up to the respective depth. $C_{2_rein_{10}}$ is a hybrid of these two, only keeping queries at level 2, but the probability estimates are derived by walks of up to 10 steps into the click graph.

Clustered Results. The $C_{2_cluster}$ method works similar to C_2 but scores are determined solely by the probabilities of moving from each query to any of the adjacent URLs. Queries at level 2 are clustered based on their co-clicked urls. Each such URL has a score based on clicks from level 2 queries. The URL score is then added to the scores of its level 2 queries. Starting from the graph formalization in Section 3, we can define the scores for a level 1 or 3 URL u_i based on the click counts from level 2 queries as $S_{url}(u_i) = \sum_j \frac{C(q_j, u_i)}{\sum_k C(q_j, u_k)}$ where j ranges over all the queries for which u_i was clicked and k ranges over all URLs connected to the query q_j . Level 2 query scores are then computed as $S_{query}(q_j) = \sum_i S_{url}(u_i)$ where u_i are all the clicked URLs for query q_j . For example, in Figure 1A, the score of $q_{2,2}$ would be a sum of the scores of its URLs, $u_{1,2}$ and $u_{3,2}$ (where $u_{1,2}$'s score is the average of clicks from $q_{2,1}$, $q_{2,2}$ and $q_{2,3}$).

Loops in the Graph. $C_{2_loop_{10}}$ differs from C_2 by keeping only queries which can be reached via multiple paths starting from the given ER query (i.e., those that are connected via URLs at deeper levels, in this case up to 10 steps). This approach would keep only $q_{2,2}$ and $q_{2,3}$ in Figure 1A. A level 2 query q_i is only considered if the path after ten steps from the origin goes through a different level 2 query and comes back to the query q_i . This approach still uses the computed probability distribution to rank entities but limits the retrieved set to those well connected in the click graph. Therefore, the queries ranked for $C_{2_loop_{10}}$ are a strict subset of those ranked for C_2 , following the same ordering.

Simple Random Walk on the Session Graph. We perform a random walk over the session graph starting from a given ER query up to 5 steps away. Please note that 1 step on the Session Graph is equivalent to 2 steps on the Click Graph, where every other step ends on a URL, rather than a query. Considering the Session Graph we compared the following approaches for ranking entities. S_5 : Starting from the original query (the ER query), walk to all queries reachable in 5 steps and rank them by their random walk probability as described in 5. Analogous, S_1 ranks all the reached queries by their random walk probability when the random walk is performed on the first level only. That is, it does not explore the session graph at queries further away than those directly connected

to the starting query. In Figure 11B, these would be the queries depicted on Level 1. Analogously to $C_2_rein_{10}$, $S_1_rein_5$ forms a hybrid method.

4.2 Combining Click Graph Results with Session Graph Results

In order to exploit the two different graphs for answering the same query we can also use data fusion approaches given the two obtained rankings. In this paper we follow the simple approach of summing retrieval status values (RSVs) used for ranking entities for each approach³ and normalizing them by the maximum score. In this way we combine scores computed with the click and session graph.

Union. As first approach, we unite the two sets of results retrieved from the click and session graphs. Their relevance scores (i.e. random walk probabilities) are normalized for each of the two approaches and if a result item appears in both result lists, these scores are added. We label these approaches as $U_{C,S}$ e.g. U_{C_2,S_1} in the case of the union of C_2 and S_1 .

Intersection. We also rank entities combining the results of the random walk on the two graphs by keeping only results which are retrieved by both approaches. Again, the relevance scores from the single approaches are normalized and then added together. Such approaches are labelled as $I_{C,S}$ e.g. I_{C_2,S_1} for intersecting results from C_2 and S_1 .

5 Evaluation

Experimental Setup. We use a query log from Bing⁴. It contains a sample of the most often clicked 35 million queries that were submitted over a period of 15 months by US American users to the search engine. This data consists of query as well as click specific details. Only query–URL pairs were retained for which at least 5 clicks were recorded overall. After some normalization of the queries there are 35 million unique queries and 44 million unique URLs. The session data consists of 25 million unique queries and a total 105 million unique query reformulations were recorded. For this purpose, we define a reformulation as two queries that were issued in the same search session within 10 minutes.

Ground Truth. In order to evaluate the proposed algorithms we constructed a benchmark for ER evaluation out of Wikipedia. As gold standard we use the “List of” pages from Wikipedia. The title of such a page is used as an ER query (e.g., “lakes in Arizona”⁵). The titles of the Wikipedia pages that are linked to from such a “List of” page are considered to be relevant results (e.g., “Canyon Lake”, “Lake Havasu”, ...). In order to use only queries that are more similar to typical Web queries in terms of length, we kept only those queries that consisted of 2 or 3 terms apart from “List of”. Thus we had 17,110 pages out of the total of 46,867 non-redirect “List of” pages. We matched these titles to queries in the

³ RSVs for ranking are the probabilities computed by the Markov Random Walk.

⁴ <http://www.bing.com/>

⁵ http://en.wikipedia.org/wiki/List_of_Arizona_lakes

Table 1. Results for finding entities using click and session graphs, averaged over the 82 ER queries in the evaluation set. Differences in MAP and R-Prec are statistically significant by means of Single Factor ANOVA. A * indicates statistical significant difference to C_2 and a + to S_1 (paired t -Test with $p \leq 0.05$).

Method	MAP	P@10	R-Prec	Queries Ranked	Relevant Entities Retrieved
C_2	0.1423	0.0959	0.0541 ⁺	489.54	8.79
S_1	0.1864	0.1026	0.1106*	78.61	6.87
$S_{1_rein_5}$	0.2011*	0.1123	0.1082*	76.37	6.63
S_5	0.0252* ⁺	0.0768 ⁺	0.0410 ⁺	2454724.54	40.92
U_{C_2,S_1}	0.1438 ⁺	0.1054	0.0792*	537.95	11.80
IC_{2,S_1}	0.2285*	0.1146	0.1283* ⁺	29.13	2.78

Table 2. Results for finding entities using click graphs. Statistical significance numbers are given to the same baselines in the previous table.

Method	MAP	P@10	R-Prec	Queries Ranked	Relevant Entities Retrieved
C_2	0.1423	0.0959	0.0541 ⁺	489.54	8.79
$C_{2_cluster}$	0.1490	0.1069*	0.0597* ⁺	489.72	8.79
$C_{2_loop_{10}}$	0.1533*	0.1077*	0.0647* ⁺	358.16	8.45
$C_{2_rein_{10}}$	0.1490	0.1069*	0.0597* ⁺	489.72	8.79
C_{10}	0.0548* ⁺	0.1	0.0549 ⁺	87313.18	35.48

log and kept the ones which appear at least 100 times in the query log and had at least 5 clicks on results. After this, we were left with 82 queries for evaluation.

Results. As a pre-processing step, all queries, both from the ground truth and from the query logs have had the stop words removed and were stemmed afterwards. We consider a retrieved entity to be relevant to an ER query if the string representing the relevant entity includes the ER query. In order to compare the different ranking approaches, we computed Mean Average Precision (MAP), precision for the first ten results (P@10) and R-Precision (R-Prec) of the produced rankings.

In Table 1 we compare our baseline runs C_2 and S_1 which are equivalent to ranking the queries directly connected to the user query by the weights on the edges. We can see that by using a Session Graph we obtain better results for ER queries. Moreover, while using the intersection of the Click and Session Graphs reduces the result set size significantly (29 results instead of 489 and 78 respectively), it improves effectiveness scores. With this simple approaches recall is anyway very low as the average number of relevant results per query is 83. The approach of unifying the sets of entities retrieved from the two graphs is not performing well mainly because of the large amount of retrieved entities.

In Table 2 we compare results of different approaches on the click graph (see Section 4.1). Our baseline is again C_2 , that is a 2-steps random walk starting

from the user query node, which is equivalent to ranking connected queries by the weights on the edges. We can see that a longer random walk (e.g., 10 steps away from the starting node, $C_2_rein_{10}$) gives a better estimation of the relevance of level 2 queries. Moreover, we see that retrieving only queries that are also supported at deeper levels in a 10-step walk (i.e., $C_2_Loop_{10}$) improves the effectiveness. Here, most of the relevant entities retrieved are kept while on average more than 100 non-relevant are discarded.

6 Conclusions

We presented approaches for answering ER queries exploiting human behaviour stored in search engine query logs. After constructing click and session graphs out of the logs, we perform a Markov random walk on the graphs in order to rank queries which contain relevant entities to a given ER query. We created a gold standard out of Wikipedia “List Of” pages which can be reused for evaluating and comparing ER algorithms. Experimental results showed that integrating results from both the click and the session graph yields best effectiveness. Such results are promising as they would allow to build systems that, given a user ER query, can answer in real time with no need of highly complex algorithms.

Future work involves developing methods for grouping retrieved queries based on different similarity measures and extracting the core representative query for each group. This way, for an entity ranking query, we can present the results to the user as a short list of query representatives.

References

1. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: SIGIR (2006)
2. Balog, K., Bron, M., de Rijke, M.: Category-based query modeling for entity search. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 319–331. Springer, Heidelberg (2010)
3. Bast, H., Chitea, A., Suchanek, F., Weber, I.: Ester: efficient search on text, entities, and relations. In: SIGIR (2007)
4. Cheng, T., Chang, K.C.-C.: Entity search engine: Towards agile best-effort information integration over the web. In: CIDR (2007)
5. Craswell, N., Szummer, M.: Random walks on the click graph. In: SIGIR (2007)
6. Cucerzan, S., White, R.W.: Query suggestion based on user landing pages. In: SIGIR (2007)
7. de Vries, A.P., Vercoustre, A.-M., Thom, J.A., Craswell, N., Lalmas, M.: Overview of the inex 2007 entity ranking track. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518, pp. 1–11. Springer, Heidelberg (2007)
8. Demartini, G., Firan, C.S., Iofciu, T., Krestel, R., Nejdl, W.: A model for ranking entities and its application to wikipedia. In: LA-WEB (2008)

9. Demartini, G., Firan, C.S., Iofciu, T., Nejdl, W.: Semantically enhanced entity ranking. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., Wang, X.S. (eds.) WISE 2008. LNCS, vol. 5175, pp. 176–188. Springer, Heidelberg (2008)
10. Guo, J., Xu, G., Cheng, X., Li, H.: Named entity recognition in query. In: SIGIR (2009)
11. Kumar, R., Tomkins, A.: A characterization of online search behavior. *IEEE Data Eng. Bull.* (2009)
12. Pehcevski, J., Vercoustre, A.-M., Thom, J.A.: Exploiting locality of wikipedia links in entity ranking. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 258–269. Springer, Heidelberg (2008)
13. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW (2007)
14. Vercoustre, A.-M., Thom, J.A., Pehcevski, J.: Entity ranking in wikipedia. In: Avanzi, R.M., Keliher, L., Sica, F. (eds.) SAC 2008. LNCS, vol. 5381, pp. 199–213. Springer, Heidelberg (2009)

Examining Group Work: Implications for the Digital Library as Sharium

Sandra Toze and Elaine G. Toms

iLab, Faculty of Management
Dalhousie University, Halifax, NS Canada
{Sandra.Toze, etoms}@dal.ca

Abstract. Digital libraries have the potential to be rich interactive environments or “shariums” that support students who work in groups to complete course work. To understand how DLs might realize this potential, the processes of a single group working on a complex project over a semester were analyzed. Findings suggest that groups perform a range of tasks including administrative, communication and information seeking and retrieval, and use multiple tools and artifacts to accomplish their work. Over the course of the work, activities shift from the individual to group illustrating the need for a complex system that intertwines public and private work space. Currently DLs provide only one tool – search – that a group might use, but do not fully support groupwork.

Keywords: collaboration, group work, design, digital library, methodology.

1 Introduction

Group work is a common element of higher education [2]; course work is regularly completed by students in small teams. Those teams use digital libraries (DLs) to find information and data to synthesize and integrate, and occasionally to create new knowledge. Marchionini [13] conceived a framework for considering the digital library (DL) as a rich interactive environment which he called a “sharium.” He proposed that the services were for use by individuals working alone or with others, or by groups who use the DL resources to achieve a variety of goals. Today’s DLs tend to provide single user access to an ever expanding group of globally located resources. The services provided by those DL applications are primarily and almost exclusively information search and retrieval. Although a chunk of Marchionini’s vision is now realized, the rich interactivity needed by groups has yet to be achieved.

In this research, we consider how students do group work to better understand how technology should support their needs with a particular emphasis on the role of the DL. To understand the problem, we isolated and followed the work of one group completing a complex course project over a semester. While the use of a single group may be considered limiting, the examination of its process illuminates the multidimensional nature of the problem illustrating that we are still a long way from the concept of the DL as sharium.

2 Prior Work

To date we have not found a systematic examination of how student group work is completed along with the types of tools that are required to support that work, nor how the collaborative learning environment is integrated with the DL. Developments tend to be at the infrastructure level (see for example, [9]), or to build digital repositories or asset management systems such as DSpace or Fedora. Nearly 10 years ago, Dong and Agogino [4] concluded that more attention was being placed on the technology than to its ultimate use and to its ability to support the learning process.

Facilitating collaboration remains a goal of DLs. But, to date, the development of collaboration technologies has been primarily the product of computer supported cooperative work (CSCW). Many technologies that are generic in nature may be deployed by DLs including awareness features, user ratings, the ability to share search histories, and to instant message (e.g., [14]).

The most significant developments to date that have the potential to impact the DL emanate from work on collaborative search technologies. This includes CIRE [19], TeamSearch [15], SearchTogether [14], CoSense [16], CoSearch [3] PlayByPlay [26] and MUSE/MUST [17]. The assumption behind these developments with very few exceptions is a “build it and they will come.” The first development to address specifically the needs of DLs was that of Twidale and colleagues [24] whose Ariadne project selectively shared the search process and facilitated group communication. Daffodil [10] a prototype enhanced search system for digital libraries also provided collaborative features including group folders and awareness.

At the same time, tools that support student tasks have emerged including annotation and note-taking. Reimer et al [18] built a prototype to aid student writing by assisting with note-taking. Adamczyk & Twidale [1] provided multidisciplinary groups with a wiki as a collaborative tool to aid their design task. The groups requested collaborative tools but found that the wiki did not easily match the demands of their work practices. Instead the students frequently chose more freeform ways to share materials; they shared photographs through Flickr.com, stored their videos on YouTube.com, and shared internet resources using del.icio.us. Notably, the wiki was used to share ideas that might help them write a report, suggesting this technology is helpful for fairly straightforward information sharing, but not for more subtle design discussions and decisions.

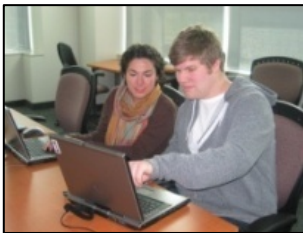
Computer supported learning environments such as Blackboard Learning Systems (BLS) provide many tools for a student and many ways of communicating with others. The emphasis is on course management, and not on facilitating the formal and informal learning with seamless integration with the DL. The DL is simply another doorway, albeit, an important one. The concept of collaboration in the context of DLs is more likely to be used with regard to the collaboration of institutions and professionals, than to support the collaboration of users.

Perhaps the most comprehensive look at how students work is Kuhlthau's [11] research on the information seeking process of students. The multiple longitudinal studies validated her Information Seeking Process model which identified core stages of the process with the associated actions, cognitions and feelings. Extending this work, Vakkari [25] observed students writing proposals, noting the close mapping of

the search process with the “work” task. One of the few to examine student groups was Limberg [12] who followed five groups of five students as they worked collaboratively on a paper over four months, but the focus was on individual learning and not on the group. Similarly, Hyldegard [8] applied both Kuhlthau’s [11] and Vakkari’s [25] models to a group environment, but the focus was on the individual as group member, and not on the group itself. Consequently, there remains a critical gap in our understanding of how a group of students work on complex tasks and how it identifies its information needs, as well as how it collects, shares and uses information.

3 Developing Methods

Part of our challenge is simply methodological – how to study groups. Much of the research has involved artificial groups with researcher assigned tasks, or the study of the individual within a group [8, 12], but not the group as a whole. To overcome the limitations and complexities involved with prior studies of groups, we developed a new approach which is explained in Toze and Toms [23].



Our method entails intertwining the natural world with the artificiality of the lab. To accommodate the need for a more naturalistic process, we studied student groups working on pre-assigned coursework over a semester. The students were not our students, and the motivation, in addition to a personal desire to learn, was the need for course credit and satisfactory or exemplary performance. Thus the task was a “real world” task, and not one constructed for a lab study. To accommodate our need for rich data that captured the interchange and interplay between and among students as well as their use of artifacts and learning tools, we opted for unobtrusive observation in a lab setting. The lab, designed for group study, contains four remotely controlled video cameras, ceiling and tabletop microphones and a set of laptops in a multi configurable

space. To use the space for group projects, the group had to agree to be video and audio recorded and to have all computer activities logged and all artifacts created by the group copied. In addition, the group added a researcher to their email discussion, and agreed to maintain diaries, and be group interviewed at the end. In turn, they could book the room according to the group’s schedule.

This method was designed to meet four key criteria that are elemental to group-work: a) time, b) motivation, c) interaction, and d) individual and group level cognition. We captured data over *time* (rather than just a single session) as much of group work occurs over multiple instances; we had *motivated* participants who worked on their tasks which had consequences for the students; we captured the *interactivity* among members; and finally, using multiple data collecting instruments we captured both individual and group *activities and contributions*. We call our method a “group naturalistic-lab study,” adopting the best of both worlds.

4 Methodology

We used this method with a group of graduate student who met for multiple sessions on a complex group project, isolated the activities that took place over the term, and identified core technological needs that are required to support their learning.

4.1 The Group

The group included four graduate students ($m=1$, $f=3$) who ranged in age from 21 to 27. They were in the first year of a Masters program in a professional-focused partially science - and partially social science-based discipline. They had not worked together, or known each other previously. Three had science and one had social science undergraduate degrees. The group formed naturally from the class from which it emanated, and was not artificially created.

4.2 Task

The group project – the task – involved five educational activities that form our “experimental” sub-tasks: 1) doing a literature review to identify topics in the conceptual framework topic; 2) identifying 2 to 3 core readings to assign to the rest of the class; 3) delivering a substantive presentation with visual aids 4) identifying and scheduling guest speakers; and finally 5) designing and leading a class exercise. The task, assigned by the course instructor, was comprehensive, and with consequences for non-completion beyond the scope of this research. The entire exercise was worth 50% of the course grade.

4.3 Procedure

In their first visit to the Lab (see Figure 1), the group was introduced to the facilities including a visit to the Control Room to view multiple displays of camera output and recording devices and to their working space in the GroupWork Lab. Part of this visit also was an orientation to our research so that they were aware that we were interested in process, and not in the specific nature of their course and class work. We wanted them to feel comfortable knowing that we were not assessing what they were actually doing.

During that visit each individual member signed a consent form, and filled out a questionnaire to provide a demographic and work experience profile. The group was also invited to ask for any tools or materials that they needed to complete their work. The researcher then left the room, and invited the group to work on its project.

For subsequent sessions, the group booked the Lab when they needed a meeting. The room was returned to the position in which the group left it, i.e., replacing any papers, charts or files left on the laptops they had left from the previous visit. The cameras, and microphones were activated, and Morae restarted on the laptops. There was no set number of sessions; the group decided when the final session was held.

Between each session, each group member completed a diary which asked them to identify and estimate time spend on the following activities related to the assignment: planning, gathering materials, writing, assessing, revising/editing/proofreading, and

other. The diary also encouraged them to reflect on specific challenges they might have faced that week, whether they shared information found or asked for help with information seeking tasks, if they learned anything from other members, and if they either used technology or identified a need for it. In addition, one researcher was added to their email discussion list.

At the end of the project, each group member filled out a survey independently, and met with one researcher for a group interview. The individual survey included questions related to information sharing practices, as well as open ended questions related to how well they thought they worked together, duplication in processes, and how they would rate the final product. The group interview asked them to collectively assess their process and results, specifically asking about their motivation, information sharing practices and if they would choose to work together again. This provided individual and group level reflections on key group processes.

4.4 Data Analysis

Four types of data were collected: unobtrusive observation via video feed; conversation via the audio feed; log files that showed how the technologies were used; diaries and email to intercept exchanges between formal meeting discussions; pre and post questionnaires completed by individuals; and a group interview. The 14.3 hours of video recording were loaded into NViVO for coding. Coding was done iteratively in stages, beginning with temporal patterns; how the group moved from one activity to another, and what tasks they were accomplishing. We identified information tasks performed, tools used or requested (e.g., flipchart, computer), artifacts created (e.g., outline, presentation) and resources accessed. Moments when the group seemed to have a shared understanding were noted, as well as points where knowledge transfer occurred. During this process 18 information searching and retrieval episodes (i.e., when a search for information was initiated, searched and resolved) were identified. The Morae logfiles were extracted and examined to provide details of who performed which task, the queries entered and reformulated, as well as specific websites, articles and information examined. The emails, diaries and interviews were examined to identify and provide details of information seeking between sessions, and to understand the individual and group assessment of how well the group worked.

5 Results

The Group met six times in sessions of two to three hours over a period of three months. While the group appeared aware, and made comments about the cameras in the first session, they quickly became comfortable in the space, and used it as their own. While completing the assignment, the group moved through the following stages: pre-focus, focus formulation, and post-focus, as defined by Vakkari [25]. A similar set of stages was observed within within each session. At the beginning of each session the group spent a few minutes in greetings, general discussion and updates, before beginning task-related work. At the end of each session there was a process of confirming what had been achieved, agreeing on new goals, and ensuring appropriate division of labour and deliverables for the next session. In our analysis,

we concentrate on the activities and artifacts used by individuals and the group in the completion of the course project. We first examine the activities that occurred within each session and then between sessions. Finally, we extract the tasks that were performed, the tools used for each task and the group artifacts created.

Session 1: In the first session the group members became attuned to the room and to each other. They moved the desks and computers to try and make the space their own, and did refer to the process of being “watched.” To help organize their work, specifically Sub-Task #3, the presentation, one member brought an outline that they had prepared in advance. This outline was to become a group artifact, and was updated and revised throughout the remaining sessions.

Some individuals brought books and, one shared a specific article while all referred to class notes. Information seeking occurred spontaneously, within the meeting, and related to discovering general background information as well as fact-finding. Originally the group decided that they did not need computers; they just wanted to discuss. But they reached for a computer when they realized that finding information was essential to their discussion. Three independent incidents of information seeking occurred during this session. The participation in the information seeking was fluid; at times an individual worked alone while at other times it was a pair activity, and occasionally the whole group participated. A key information seeking moment involved searching for a video of a media release that announced a key change in the area that they were investigating. While information was viewed and discussed the group did not make any major decisions, and commented on being overwhelmed by their topic and the perceived information needs.

At this first meeting some technology related needs were identified, which persisted throughout the remaining sessions. The group specifically discussed issues related to saving and sharing information resources such as articles, books, and websites. They discussed whether their class BLS site, could be used as a common project space, where they could store the relevant information resources as they were found. They were not sure how to make this work but concluded that as long as they were consistently sharing new resources they should avoid duplicating work.

The meeting ended with a division of tasks, each member taking on an area of interest from the overall topic. They agreed that between meetings they would all do independent work on their topic and bring a summary to the next meeting. The group also agreed that by the next meeting they should alert each other of necessary changes to the outline based on individual reading, and of any potential overlaps.

Session 2: After the initial meet and greet and updates, session two involved a focus on task #3, the group presentation. The outline was used as a key group artifact, and they collectively tried to update and revise the document. At this point, the group was still assessing the scope of the topic; they were concerned about the boundaries between their presentation and other topics within their class. Knowledge sharing from the individual to the group was a frequent activity, as the group questioned, clarified and confirmed information shared.

During this session, the group identified eight information needs and subsequently searched to resolve those needs during this session, some of which were information needs nested within another information need, as they reflected on the found

information. Participation in these activities varied from individual to pair and group; there was frequent shared computer use. The group searched for information on a government association, and regulatory information; one member brought an article to share. They used this information to help make decisions, to confirm items in their outline or add new ones. During the information seeking, sometimes individual members worked privately to search and shared only particular findings, while occasionally the group participated.

The group expressed a need to use information collectively. They asked for a flip chart and used it to reproduce their outline. In addition the group expressed a need for a space to keep “great ideas” or key thoughts. Members were reminding each other to write things down and to keep track of critical items that they needed to revisit.

The session ended with the sense they had started off overwhelmed and worried, but that had made good progress during the meeting. They would continue between meetings to flesh out their individual sections of the outline.

Session 3: After initial greetings and non-task related chatter, the group began working on tasks. In this session they branched out to discuss Sub-Task #4, the identification of guest speakers and Sub-Task #5, the class activity as well as Sub-Task #3. In course of this work, they used books, websites, and articles, as well as phone calls to experts. They identified four information needs which included fact finding as well as information needed to clarify the scope of their topic. Participation again moved from individual activities to pair and group discussions. While no new technology needs were identified, the group continued to struggle with sharing critical information. In particular, they discussed how citations should be included in the presentation to ensure consistency and standardization.

In addition to continuing to work with the outline the group concluded that a flow chart would also help them visualize their topic together. The flip chart was used for this with one group member consistently writing, and others searching to solve questions that emerged. The session ended with the sense that the flow chart was not quite accurate, but that they were moving in the right direction. They emphasized the importance of collectively pooling information into the flow chart. One participant remarked “this is how we are going to get through this.”

Session 4: During this meeting the group spent much of this session on Sub-Task #2, choosing core articles; Sub-Task #4, the identification of a guest speaker; and Sub-Task #5, the class activity. They felt the pressure of time and the need to make a decision on the location for the group activity and the number and potential topics for their guest speakers. In the meantime, they continued to work on the flow chart and outline related to Sub-Task #3. To help make decisions, searching was initiated to determine directions and locations and articles were sought to confirm selections for the class. This information seeking was intertwined with the various tasks and completed by individuals and pairs; the results were discussed by the group.

For example, a book was identified as critical to their topic, and all members were urged to read the relevant sections. The group discussed, and negotiated and integrated information created by individuals, often by specific reference to individual notes, core articles and books. No new technology needs were identified, but the group used parallel processes to ensure they could see the information together both

inside and outside of the meeting. The flow chart that they were creating on the flip chart was transferred to a publishing software package to create a portable version.

Sessions 5 and 6: Session #4 was the last session that included information seeking. After Session #4, the emphasis moved from collectively examining and assessing information to working with what they had gathered, and combining individual sections into their final draft of the group presentation for Sub-Task #3. In Session #5, they discussed and identified a style and template for their digital presentation, while in Session #6 they spend much of the time integrating each member's slides. This process took a long time, and members seemed frustrated with the process. The group requested and received a projector for these last two sessions so they could practice the presentation. This process did not include all of the prepared slide deck, resulting in some conflict between individuals as they tried to ensure that a coherent "group" focus was apparent throughout. They questioned and confirmed the information, the sources, the order and even how language was used by specific members. The sessions ended due to time pressures, and without a sense that the presentation was complete. They were to individually examine and review slides given the whole presentation, and would quickly run through things before class.

Between Sessions and at the End: Individuals did substantive work on their sections between meetings, while meetings involved integrating, synthesizing and making decisions. The email and diaries show that members were individually searching for information between meetings. They exchanged emails to update members on their communication with potential guest speakers, and to alert others to important information resources. One member in particular sent links to articles, and recommended books. Others also sent members recommended articles. All members updated the outline with new information based on their readings, and circulated the draft to the others.

The final interview with the group revealed additional aspects of group process. While all members agreed that they enjoyed working together, and thought that overall their processes were fairly efficient, several members thought the group had problems with the division of labour and that the final product lacked flow. One member was quite critical suggesting there had been too much individual work and not enough integration at the group level. This individual attributed this partially to the group itself, but also to the topic which was challenging, with much overlap between the subtopics. In terms of information seeking, the need to identify and share critical sources earlier in the process was recognized. For example, the book that was discussed in Session #4 was determined to be a critical resource, and that their work could have been more directed had it been found earlier. The inefficiencies in sharing information, and the divergence in work habits and timing of individual work were also noted. The group identified the flip chart and the projector as useful for helping them work as a group.

Tasks and Artifacts

The group completed a series of tasks that can be categorized as:

a) administration: scheduling or logistics, determining formats for slides, timing for the different parts of the presentation, obtaining documentation from guest speakers, and directions and plans for the class activity,

- b) communication: updating the group on individual activities; contacting potential guest speakers, using email to share information sources and materials,
- c) problem solving: discussing aspects of the topic, negotiating positions, making decisions about readings to assign to the class, guest speakers and the organization of individual sections for the presentation.

To complete these tasks the groups worked with information in the following ways:

- a) finding support material: completing 18 instances of information searching and retrieval [23] that occurred within the sessions, and sharing of articles, books, web-sites and names of key experts between meetings,
- b) creating new artifacts: developing and amending the task outline, creating flow charts, and determining the focus and flow of their presentation.

They used multiple tools that included:

- a) a flipchart to create and recreate the outline and flow chart,
- b) presentation software to create individual and the group presentation,
- c) search engine and databases to search for a range of information including background information on government agencies, associations and legislation, key articles, directions, and to confirm facts including names, dates and people,
- d) documents including books and articles (digital and print),
- e) a projector to practice the presentation as a group,
- f) the computer which was generally used as personal space, but sometimes as common space, as members examine an individual computer display together.

They created multiple artifacts to support interim steps, and the final products:

- a) an outline of the presentation,
- b) a “bucket” to keep important facts, decisions and thoughts,
- c) a flow chart on paper and in electronic form to manage their topic,
- d) a group presentation that combined sets of individual slides.

6 Discussion

The inner workings of this group demonstrate the complexity of what at the onset appeared to be a simple and typical course assignment. Multiple sub-tasks were required to complete the work using multiple types of tools and multiple types of information resources. At the same time the group created different types of artifacts to support the meeting of those goals.

6.1 Group Process

Over the course of six sessions, this group performed a range of tasks that included routine administrative tasks to organize the group and its process, and a series of communication tasks to facilitate exchange among members when collocated, and when working remotely. These activities are what would be expected of a group, and represent many of the generic applications that groupware tends to provide (see [19]).

The results additionally illustrate how information seeking and retrieval are tightly interwoven into the work task process as previously noted by [5,8]. The process of seeking information was triggered by some aspect of the group work that filled a gap in knowledge or fulfilled a desire to learn [22]. The source of that information varied

from physical books to articles found through the University's DL and information found via web search engines, as well as from human sources.

At the same time, multiple information objects were created. The group developed an outline, first in paper, and later in digital form to sketch their thinking about how the topic matter was unfolding. A flowchart was created to visualize how the levels of agencies connected. The group using multiple individual notes to keep track of information noting a need for a "bucket", to ensure that information was not lost.

Over the course of each session, it became clear that the work of the group shifted back and forth from the individual to the group, as also noted in [8]. Even during group meetings, the notion of public and private spaces was evident, as sometimes a member would work privately on an aspect before revealing a result to the group. Except for the projector and flipchart, all of the tools used by the group were individual such as the computer and telephone. Sometimes the individual tool needed to be a group tool such as when the group crowded around a single display, but this tended to happen toward the end when the group collated materials for final products. That said, having a common storage space was important. This served many purposes some formal, such as a common place for references used throughout the lifecycle of the project, and some less formal such as a place to record thoughts and ideas that might prove useful. At the same time, group members needed to be aware and made aware of individual work completed between meetings.

6.2 Supporting Collaboration in the Digital Library

The DL as currently rendered by many organizations is just another tool and not necessarily *the* tool of choice for information search and retrieval by students. At one time, the library catalogue provided access to its resources; today it tends to provide access to global resources. For the DL to embrace the sharium concept, it must augment the services that it provides and enable richer forms of interactivity.

As our research and that of others [5,7,25] show, it is not just about the search and retrieval process, but very much about how the need emerges from the work process, how found information is subsequently integrated and used, and how new knowledge – new artifacts are created. A DL needs to support these dynamic and fluid processes.

From our analysis, it seems that there is a combination of public and private needs. On the one hand, members explicitly conversed and interacted with each other verbally and through sharing physical artifacts and jointly examining the contents of computer displays; on the other, they pondered, and retained a personal set of notes, ideas, and documents, and waited for the "right" moment to share with the group. This suggests the need to provide public and private spaces not unlike the findings of [20] in their study of collaborative tabletop use.

At times the group subdivided into pairs, or singles or an individual plus threesome depending on the nature of the task. Sometimes the group clustered around a single computer or the flipchart, while at other times, they independently worked on artifacts. They clearly needed multiple types of shared space for maintaining ideas, and progress, as well as citations. The bookbag concept in wikiSearch [21] is a starting point. But the challenge is in a seamless interface that helps a group to manage its process and progress, while enabling independent activities, from finding appropriate pieces of information to building interim and final information objects, and respecting a member's need for private workspace as well as shared workspace.

7 Conclusions

Although Marchionini [13] proposed the concept of the sharium a decade ago, the notion of collaboration and how it could be managed within the DL has not yet been achieved. There is, however, considerable work in collaborative tools within the CSCW community (see for example, [14]), but that has yet to be integrated with the DL. In higher education, computer supported learning environment such as Blackboard Learning Systems administer courses and facilitate communication rather than learning. Within these, the DL is but a link within the interface.

Group work is a common activity in educational settings as well as in the workplace; we now need to consider innovative ways of integrating the multifaceted groupwork process that represent a fusion of individual and group activities into DL services. While our data is based on a single group, that single group demonstrates the complexity of the problem. The DL was only one tool, one search engine from among a host of other tools and information access products used by the group. Our research shows that even a class assignment completed by a group is a complex problem with multiple interconnected subtasks that need multiple tools and create many artifacts. What is the role of a DL within that larger work process? Is it merely a search and retrieve tool, or should it augment its services to seamlessly manage all of the various information processes?

Acknowledgments. This research was supported by grants to Toms from the Canada Foundation for Innovation, the Canada Research Chairs Program, the Natural Sciences and Engineering Council of Canada NECTAR Strategic Network grant. The authors gratefully acknowledge the assistance of Tayze Mackenzie, Alexandra MacNutt, Lori McCay-Peet and Janet Music.

References

1. Adamczyk, P.D., Twidale, M.B.: Supporting Multidisciplinary Collaboration: Requirements from Novel HCI Education. In: CHI 2007, ACM, San Jose (2007)
2. Akkerman, S., et al.: Reconsidering group cognition: From conceptual confusion to a boundary area between cognitive and socio-cultural perspectives? *Educational Research Review* 2(1), 39–63 (2007)
3. Amershi, S., Morris, M.R.: CoSearch: a system for co-located collaborative web search. In: CHI 2008. ACM, Florence (2008)
4. Dong, A., Agogino, A.: Design principles for the information architecture of a SMET education digital library. In: JCDL 2001, June 24–28, pp. 314–321 (2001)
5. Fidel, R., Mark Pejtersen, A., Cleal, B., Bruce, H.: A multidimensional approach to the study of human-information interaction: A case study of collaborative information retrieval. *JASIST* 55, 939–953 (2004)
6. Grudin, J., Poltrock, S.E.: Computer-Supported Cooperative Work and Groupware. *Advances in Computers* 45, 269–320 (1997)
7. Hansen, P., Järvelin, K.: Collaborative information retrieval in an information-intensive domain. *Information Processing and Management* 41, 1101–1119 (2005)
8. Hyldegård, J.: Beyond the search process - Exploring group members' information behavior in context. *Information Processing & Management* 45, 142–158 (2009)

9. Iverson, L.: Collaboration in digital libraries: a conceptual framework. In: JCDL 2004, Tucson, Arizona, June 7-11, p. 380 (2004)
10. Kriewel, S., Klas, C.P., Schaefer, A., Fuhr, N.: DAFFODIL- Strategic Support for User-Oriented Access to Heterogeneous Digital Libraries. *D-Lib Magazine* 10 (2004)
11. Kuhlthau, C.C.: Seeking Meaning: a Process Approach to Library and Information Services. Libraries Unlimited, London (2004)
12. Limberg, L.: Three conceptions of information seeking and use. In: Exploring the Contexts of Information Behaviour, pp. 116–135. Taylor Graham Publishing, London (1999)
13. Marchionini, G.: Augmenting library services: toward the sharium. In: International Symposium on DL, Tsukuba, September 28-29, pp. 40–47 (1999)
14. Morris, M.R., Horvitz, E.: SearchTogether: an interface for collaborative web search. In: User Interface Software and Technology. ACM, Newport (2007)
15. Morris, M.R., Paepcke, A., Winograd, T.: TeamSearch: comparing techniques for co-present collaborative search of digital media. *TableTop* (2006)
16. Paul, S.A., Morris, M.R.: CoSense: enhancing sensemaking for collaborative web search. In: Human Factors in Computing Systems. ACM, Boston (2009)
17. Reddy, M.C., Jansen, B.J., Krishnappa, R.: The role of communication in collaborative information searching. *ASIST* 45(1), 1–10 (2008)
18. Reimer, Y.J., Brimhall, E., Sherve, L.: A study of student notetaking and software design implications. In: WBE 2006, Puerto Vallarta, Mexico, January 23-25, pp. 189–195 (2006)
19. Romano, J.N.C., Roussinov, D., Nunamaker, J., Jay, F., Chen, H.: Collaborative information retrieval environment: Integration of information retrieval with group support systems. In: 32nd Hawaii International Conference on System Sciences, pp. 1–10 (1999)
20. Scott, S.D., Carpendale, M.S., Inkpen, K.M.: Territoriality in collaborative tabletop workspaces. In: CSCW. ACM, Chicago (2004)
21. Toms, E.G., McCay-Peet, L., Mackenzie, M.T.: WikiSearch – From Access to Use. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 27–38. Springer, Heidelberg (2009)
22. Toze, S., McCay-Peet, L., Toms, E.G.: Information seeking and retrieval by group (in submission, 2010)
23. Toze, S., Toms, E.G.: Methodological issues in the study of group information processes (2010) (in submission)
24. Twidale, M.B., Nichols, D.M., Paice, C.D.: Browsing is a collaborative process. *Information Processing & Management* 33(6), 761–783 (1997)
25. Vakkari, P.: A theory of the task-based information retrieval process: A summary and generalization of a longitudinal study. *Journal of Documentation* 57(1), 44–60 (2001)
26. Wiltse, H., Nichols, J.: PlayByPlay: collaborative web browsing for desktop and mobile devices. In: Human Factors in Computing Systems. ACM, Boston (2009)

Architecture for a Collaborative Research Environment Based on Reading List Sharing

Gabriella Kazai¹, Paolo Manghi², Katerina Iatropoulou³, Tim Haughton¹, Marko Mikulicic², Antonis Lempesis³, Natasa Milic-Frayling¹, and Natalia Manola³

¹ Microsoft Research, Cambridge, UK

² Istituto di Scienza e Tecnologie dell'Informazione,
Centro Nazionale delle Ricerche, Pisa, Italy

³ Department of Informatics, National and Kapodistrian University of Athens, Greece
v-gabkaz@microsoft.com

Abstract. Scholarly research involves a systematic study of information sources in order to establish facts and reach new conclusions. It encompasses survey, analysis, evaluation, and creation as distinct phases that are performed iteratively and often in parallel by accessing a range of local and remote resources. Throughout these activities scholars create collections of relevant work, ranging from publication references to new information acquired through experiments or correspondence with other scholars. We use the term *reading list* to refer to such collections. Existing software packages or web services for managing publication lists, like CiteULike, lack integration with researchers' workflow which may require access to both desktop and online resources. In this paper we describe the architecture and system design of ScholarLynk, a desktop tagging tool that enables researchers to build and maintain reading lists across distributed data stores, in collaboration with other researchers.

Keywords: Desktop tagging tool, scholarly research, reading lists.

1 Introduction

Scholarly research involves a systematic investigation and study of materials and information sources to establish facts and reach new conclusions [3]. Researchers conduct such work by searching and reading relevant publications and by communicating and sharing knowledge with their peers. As a result they compile and maintain valuable collections of references and resources, here referred to as *reading lists*, each representing a body of work in a particular area of scholarly interest.

With the widespread proliferation of online repositories and specialized search engines such as Google Scholar, researchers increasingly manage their publication references electronically, using reference management systems such as BibTeX, End-Note or RefWorks. Recently, support for this type of activity is also provided by online services like CiteULike, Mendeley, Zotero, and Connotea. These sites incorporate social networking features and foster informal communication and sharing which can increase research productivity. While valuable, these Web-based services are

separated from the research activities that are carried out within the desktop environment as part of the researchers' workflow.

A typical research workflow involves several phases, including the *survey*, *analysis*, *evaluation* and *creation* of resources [9]. Depending on the nature of the task and the work style, such activities may occur iteratively and in any combination or order. Each activity can be multifaceted, involving communication with other scholars and using both local resources on the researcher's desktop and remote resources on the Internet. Thus, to accomplish their tasks, researchers continuously switch between local and remote resources and applications, often carrying the burden of coordinating and synchronizing the two in a consistent way.

In this paper we describe the design and architecture of ScholarLynk, a system that aims to support researchers in building and maintaining reading lists in collaboration with others. Central to the design is the generalized concept of a reading list which we extend to designate a grouping of any types of resources, from academic papers to web pages and emails, and to include comprehensive metadata, annotations, and associations among items in the list. The novelty of ScholarLynk is twofold: (i) it implements tagging in the desktop environment to enable the collection of resources across distributed storage locations by an individual or a group, thus providing a unified interface for managing desktop and web data sources, and (ii) it provides an architecture that unifies search over different content repositories, tagging, and social interaction, thus supporting a collaborative environment.

The paper is structured as follows. We first reflect on related work (Section 2), then present several use cases that motivated system requirements and designs for ScholarLynk (Section 3). In Section 4, we describe ScholarLynk's architecture and implementation. We close with conclusions in Section 5.

2 Related Work

In this section we first describe the scholarly research process that we aim to support in ScholarLynk. Then we review research on online and desktop tagging practices and solutions for organizing resources, such as academic papers.

2.1 Scholarly Research Process

Kolb [9] defines the scholarly process as a methodology for making references to ongoing debates from a series of sources and arguing for claims. He suggests that this process typically encompasses four phases: survey, analysis, evaluation and creation. The activities that comprise these phases include searching (e.g., direct searching, browsing, chaining, monitoring), collecting (e.g., gathering, organizing), reading (e.g., scanning, extracting, assessing, re-reading), writing (e.g., assembling, disseminating), and collaborating (e.g., consulting, coordinating, networking) [1,2,3,13]. All these activities feed into one another, creating an intertwined, iterative, and often parallel research workflow. For example, as researchers search for information they continually alter their information needs through monitoring, differentiating, and extracting new knowledge [5], leading to further searches. In addition to search, scholars often locate further readings by chaining, i.e., following connections from a useful paper to other sources, e.g., by reviewing cited works or other papers written by the author [2],

or by browsing semantically similar articles [1]. Finding links between research topics is often done through informal communication, e.g., email, meetings, or collaboration between scholars [13].

As a result of their investigations, scholars build personal collections that support their current and long-term research [4,6,8]. Items of interest may include previously written papers or papers published by others, notes, e-mail communications with peers, and web pages [12]. The process of making sense of the gathered information involves posing questions, making hypotheses, and expressing opinions on what the collected pieces are and how they relate to one another. Understanding emerges through re-visitation and by constructing a coherent story and structure toward a publishable form [11].

In ScholarLynk we aim to provide a framework for supporting the scholarly process in a unifying manner, respecting the required flexibility and connection among all the stages of research. We aim to augment the desktop environment with a layer that bridges disconnected parts of researchers' existing infrastructure, instead of introducing yet another application silo. Our approach is to provide a unified view and control over resources required to conduct research. Thus, ScholarLynk enables users to use tagging across desktop and remote environments to create reading lists of heterogeneous local and remote resources and supports in-context communication and collaboration around shared reading lists.

2.2 Desktop and Online Tagging Solutions

The popularity of resource tagging has increased with the appearance of social media in online communities [12]. Users are encouraged to add metadata in the form of keywords, either from a fixed vocabulary, as in Yahoo! Answers, or by generating their own tags, as in Flickr. The benefits of such collective efforts include improved search, spam detection, and personal organization. Fundamentally, this is enabled by the underlying social structure and user interactions supported by these systems [10].

An example of a collaborative tagging system, designed specifically for the needs of scientists and scholars, is CiteULike [6,8]. Similarly to other social bookmarking sites, CiteULike allows users to tag URIs with personal metadata. Tagged resources and their metadata are then available and can be shared from a single place, i.e., a dedicated Web service. Other online tools for organizing citations include Zotero and Mendeley. Zotero is an extension for the Firefox browser that can recognize and extract data and metadata from a range of different digital libraries. Users can bookmark publications, and then add their own personal tags and notes. Mendeley is a similar application, which has both a web browser version and a desktop client [7].

A disadvantage of these online tagging systems is that the tags are only used inside their own respective repositories, creating multiple vertical worlds that users have to manage in parallel. In addition, each such system requires users to invest time and effort to learn how to use and import or enter bibliographic information.

Tagging in the desktop environment has been used in the context of activity management, enabling users to organize resources around activities [12,14]. By grouping resources into activities, users can switch between tasks, rather than between application windows, and they can access related resources from the context of a

given activity. Unlike online social media tagging, desktop tagging is motivated primarily by personal needs, i.e., to optimize access to resources.

Following the desktop tagging approach in [12] that combines disparate data and bridges application silos, ScholarLynk presents a unified view over the user's resources and workspaces. Moreover, ScholarLynk extends the notion of a tag to include richer descriptors of items in a reading list and capture semantics of their inter-relationships. Similarly to Mendeley, ScholarLynk comprises a desktop client but, rather than focusing on the desktop application itself, it provides an integrated cross-platform view, facilitating seamless management of tagged resources regardless of the service, application, or storage location. Furthermore, ScholarLynk aims to integrate various communication channels such as Twitter and Facebook in order to exploit social interactions to aid users in the discovery of relevant resources, exchange of knowledge, and collaboration around shared reading lists.

3 Use Case Scenarios

In this section, we detail a sample of user scenarios based on the literature about scholarly research, e.g., [13] outlining the shortcomings of current solutions for supporting scholarly work and highlighting the benefits of the ScholarLynk design. We base our scenarios on two personas, Zoe and Shaelyn, both Ph.D. students in Computer Science. They differ only in the way they organize their citations: Zoe uses Zotero and CiteULike, while Shaelyn uses ScholarLynk. They are both active Twitter and Facebook users, regularly staying in touch with colleagues they collaborate with. They both have existing collections of references and resources relating to their work, including collections on two specific topics: "Folksonomy" and "Crowdsourcing".

The ScholarLynk profile for Shaelyn includes the following information:

- (a) Personal data such as name (mandatory), affiliation, and own publications (stored as a reading list that is manually editable and automatically updated from selected online repositories, e.g., DRIVER and DBLP);
- (b) Network data such as links to co-authors, cited and citing authors, colleagues, etc. Included in her network are users whom she *follows*. Shaelyn also follows several reading lists published by others as well as her own publications. She receives alerts when others use or change items that she follows, e.g., when someone cites her work or when changes are made to a reading list or a user profile;
- (c) Reading lists that Shaelyn maintains are groupings of heterogeneous resources that are relevant to a research theme or an activity of her work. For example, her "Crowdsourcing" reading list includes academic papers, web pages, emails, local files, and notes.

Keeping up to date. Zoe regularly monitors her email and checks CiteULike, Zotero, Facebook and Twitter for relevant activities. She has to do this as separate tasks, logging into her different accounts. She currently has no means of tracking if and when her own publications are being cited by other researchers.

In contrast, to keep up to date, Shaelyn monitors her ScholarLynk message console, see **Fig 1**. She regularly receives messages from her peers and automatic alerts about users and reading lists that she follows. Her current messages include: (1) a note

from a colleague, recommending a paper for her Folksonomy reading list (from Twitter), and alerts that (2) a user copied items from her “Crowdsourcing” reading list into their reading list on “Social data”, (3) a colleague added one of Shaelyn’s publications to their reading list on “Social search”, (4) a new publication is citing one of Shaelyn’s papers (from DBLP), and (5) a user rated one of Shaelyn’s papers with four stars in the context of their reading list on “SocialMedia”.

Shaelyn clicks the fourth message. She is directed to a digital library and skims through the copy of the publication. She finds it relevant to her work on crowdsourcing and adds it to her reading list. After checking the ScholarLynk profiles for the authors she also decides to follow one of them.

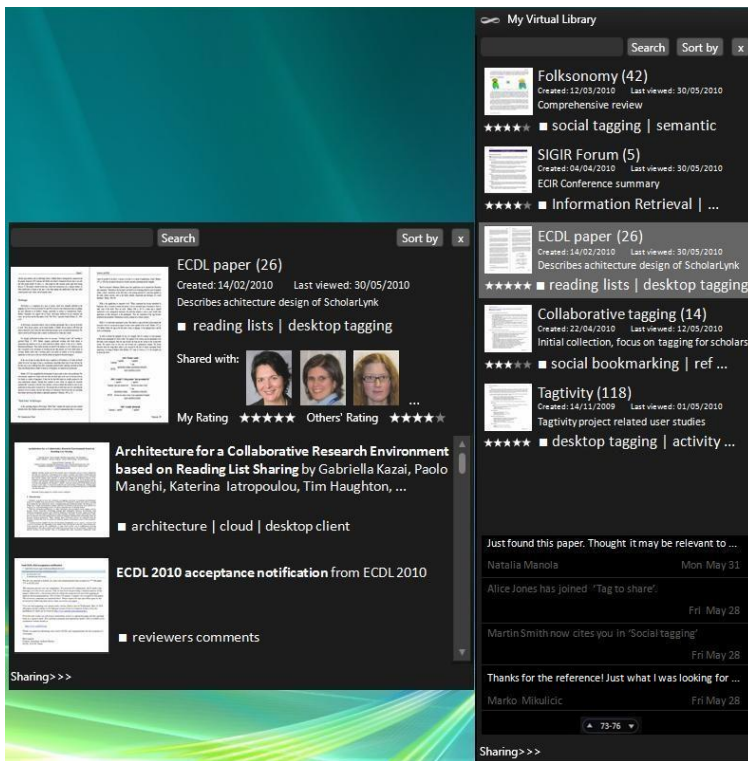


Fig. 1. The ScholarLynk desktop toolbar provides a central access-point for the user’s reading lists and resources. It also shows the user’s message console displaying incoming automatic alerts and user messages.

Search and creation of collections. Zoe wants to learn about ‘game theory’ to assess how relevant this area may be to her work on crowdsourcing. She opens a web browser and enters a query in the search box. Among the results she finds a Wikipedia article on game theory. She clicks the Zotero icon on her browser to add the article to her Zotero Library. She follows one of the citations on the Wikipedia page that takes her to a digital library that contains the cited paper. She skims the paper and

adds it to her Library too. A copy of the paper is downloaded into her Library along with the bibliographic metadata. She switches to her Zotero Library browser window, creates a new collection, and adds the two new items using drag-and-drop from the generic area of her Library.

Shaelyn is completing a similar task using Scholarly. She enters a query on game theory into the ScholarLynk toolbar. ScholarLynk searches Bing, Google Scholar, DRIVER, and CiteULike in parallel, showing the results grouped by the search providers in a browser window. In addition, it enriches the results with data and statistics from social usage of the ScholarLynk community, displaying the reading lists that contain a given search result and ‘approval’ statistics such as the weighted average of user ratings. Shaelyn hovers with the mouse over one of the reading lists associated with a search result. A popup window shows the contents of the reading list. She drags and drops a couple of the items onto her ScholarLynk toolbar creating a new reading list. As a result of her actions, an alert is also sent to the owner of the reading list, informing that Shaelyn copied items from it. At the same time, alerts are also sent to anyone following Shaelyn or the topic of game theory about Shaelyn’s new reading list.

Sharing and annotations. Zoe switches to her web browser and checks her Twitter account. She sees that one of her peers, whom she follows, tweeted about an article on folksonomy, citing the title and the authors. Zoe copies these into Google Scholar search and clicks on the search result that has a copy of the paper and skims through the article. Using the Zotero plug-in, she downloads it into her Library. She then switches to her Library and moves the paper into her “Folksonomy” collection. She jots down a couple of notes about the article in Zotero and adds tags. Zoe thinks that the article will be of interest to her team members too. She opens her group’s Facebook page and leaves a message on the group’s wall, manually including the URL for the paper in the digital library.

Shaelyn, on the other hand, receives the recommendation about the article on her ScholarLynk console. She drag-and-drops the title and author information from her message console to the ScholarLynk search box. From the search results, she drags and drops the paper to her “Folksonomy” reading list. She adds a few notes to the paper, tags it, rates it, and links it to another paper in the reading list. She then decides to share a part of her reading list with the “Social Web” group that she is a member of. She drags and drops items from her reading list to the group’s reading list. Group members are automatically alerted.

Collaboration. Zoe needs help with finding additional relevant items for her “Crowdsourcing” collection. She asks some of her colleagues via email, pointing them to her CiteULike collection that she compiled so far. Any recommendations she receives, she reviews and manually incorporates into her collection.

Shaelyn also decides to ask for help in populating her “Crowdsourcing” reading list. She writes a description of her perspective on the topic and opens her reading list to the community, flagging it as public and posting a message on her wall that she is seeking help. All users who are following Shaelyn or topics related to crowdsourcing receive an alert that she needs their help. Soon, close colleagues and a broader user community start adding to Shaelyn’s public reading list. She reviews these regularly, rating some items, removing others, and including additional resources herself.

4 ScholarLynk

Zoe's scenario above reveals the syncopated character of operations that a researcher needs to perform to exploit the rich but dispersed scholarly and social environments. Elements of her knowledge organization and collaborations survive replicated among disconnected environments on the desktop and in the different web tools, requiring additional effort to manage and control. In contrast, Shaelyn's scenario illustrates how an integrated and interconnected research environment may support researchers' workflows. *ScholarLynk* aims to provide such an environment, supporting users in their daily research activities throughout the entire lifecycle of their work and facilitating in-context communication and collaboration. This section presents the architectural design of ScholarLynk, starting with the functional requirements and the data model.

4.1 Functional Requirements

Based on our analysis of the scholarly work practices reported in the literature (see Section 2.1) and our use case scenarios outlined in Section 3, we identified the following personal and collaborative requirements.

Personal requirements. Researchers need a personal space, here referred to as *My Virtual Library (MVL)*, comprising heterogeneous *items* that are collected from the web or locally, possibly with the help of colleagues. Researchers need to be able to organize these items into *reading lists* according to a common topic, a theme of research, or temporary or long-term activity. It is important that reading lists support flexible structures where an item may belong to several lists and a reading list may refer to and thus subsume other reading lists. MVL can be considered as a "root" reading list. The types of items include publication files (e.g., PDF, DOC), email, web pages, images, audio/video, and reading lists.

Requirements to support sense-making. Researchers' work involves discovering the meaning and relationships among items in the reading lists and their relevance to the topic under investigation. They often record their *interpretations* of the items in the context of the particular topic [11]. These records can take the form of *tags*, *notes/comments* and *ratings* or simply indicate unlabelled links between items [12].

Personal workflow requirements. Researchers need to be able to manage their reading lists (create, delete, rate, etc.) and items (add, remove, rate and annotate in the context of a reading list, etc.) *as they engage in their usual research activities of survey, analysis, evaluation and creation on the desktop* [4,7]. It is important that MVL seamlessly and flexibly integrates within the researchers' existing practices, augmenting them with useful functionality without imposing yet another application-centric solution, interfering with researchers' habits, or replacing existing tools [12].

Collaborative requirements. Researchers communicate with colleagues to exchange knowledge and they collaborate with each other [13]. For that they need to share items, reading lists, and their insights and interpretations. They may provide feedback or express opinions. They may wish to keep up to date with developments in the field and evolving interests of their colleagues. Thus, they need to monitor the shared reading lists of their colleagues and broader communities of interest. They need to be notified of updates to reading lists and related activities of their colleagues and communities.

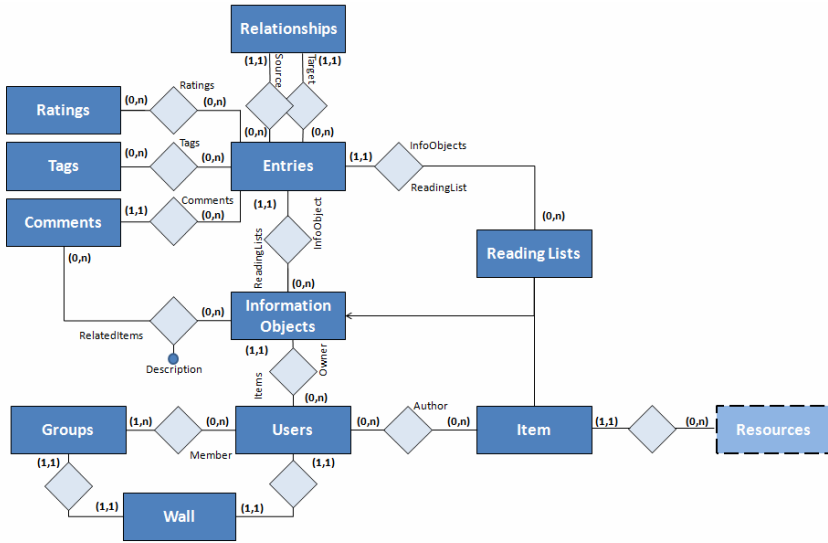


Fig. 2. ScholarLynk data model: Entity-Relationship graph

Collaborative workflow requirements. Researchers need to be able to manage their shared content (i.e., control access to items, reading lists, and elements of interpretation), and their interactions with colleagues (i.e., notifications) *in a lightweight manner that is integrated with all the phases of their usual research involving survey, analysis, evaluation and creation on the desktop.*

4.2 Data Model

The entities underlying ScholarLynk are illustrated in the data model in **Fig. 2**. Central to the model are the interrelated notions of resource, information object, item, entry and reading list.

- *Resources* are digital artifacts or real world entities which have associated digital representation or metadata. Resources may be local to the user's desktop or remotely available through the Internet. Examples include emails, academic papers, web pages, files, directories, books, people (authors/users), images, audio/video. A resource resides outside of the system and is uniquely identified by a reference, e.g., file system path, URL, DOI, or MAPI address (for email).
- *Information objects* are entities representing first citizens of the environment. They have an owner, i.e., the user. Both reading lists and items are information objects.
- *Items* are system representations of resources. They contain metadata (i.e., properties) that describe the resource, e.g., title, date of creation, thumbnail, etc. Further metadata can be provided depending on the type of resource represented by the item. For example, a publication record may be described by its bibliographic record (author, publisher, publication date, and abstract) while an email metadata may include its recipients, sender, subject, and send-date. Item subclasses can be extended to describe new types of resources.

- *Reading lists* represent an ordered or unordered collection of information objects, with metadata that captures the specific user’s interpretations. A reading list inherently defines a “context” in which the user works and to which the collected items are related to. It reflects not just the user’s view and knowledge of the topic but their current status of work. A reading list is thus modeled as a set (or list) of *in-context* Information Objects or *entries* each representing: (i) the link to the information object itself and (ii) the *interpretation* of the information object in the context of the reading list, e.g., tags, comments, ratings. A reading list is also an information object in itself. As such, it has an owner, i.e., the user who created it and can be linked to entries of other reading lists.
- *Entries* represent the participation of information objects in the reading lists. They are a combination of the item and the interpretation of the item in the context of the user’s reading list. The interpretation layer includes labeled or unlabeled *relationships, comments, tags, and ratings*.
- *Relationships* are directed associations between two entry entities, i.e., the “source” and the “target” entities. A relationship may have a label describing the nature of the connection it represents. It can reflect implicit relationships, e.g., citations between academic papers, or user-defined relationships, e.g., reading order or relevance ranking.
- *Users* may be both producers and consumers of information objects, e.g., author, owner, or follower of the content. They can form *groups* and communicate with others via public message consoles, e.g., *walls* (similarly to Facebook) or write comments in the context of information objects.

4.3 Architecture

ScholarLynk is designed to meet the personal and collaborative requirements above. Researchers can seamlessly interact with the “outside” world, searching and fetching new resources or exchanging their information objects (i.e., items, reading lists) and interpretations with other researchers. A unified view of all these actions is provided by the *My Virtual Library* (MVL) concept. In MVL researchers can access and organize all local and remote resources as reading lists of information objects, along with their relative interpretations. Uniform treatment of all supported resource types, regardless of location, is permitted by the referencing mechanisms whereby entities can be identified by their type and unique address, e.g., URL, file path, MAPI address, etc.

To support sharing, ScholarLynk is designed as a peer-to-peer architecture, with two logical layers: an overlay network, called *ScholarLynk Cloud*, to which *ScholarLynk clients* can dynamically connect to or leave, see **Fig. 3**. The Cloud handles *users*, i.e., the researchers, and their (private or shared) information objects. In particular, users are the owners of the items they create and must authenticate to share their items or be authorized to access somebody else’s items.

ScholarLynk Clients. The client software is closely integrated with the desktop experience and provides researchers with: (i) *Integration of web data sources* (e.g., digital libraries, Google Scholar, etc.) as Windows network drives which can be searched and browsed, and from which resources can be dragged and dropped into

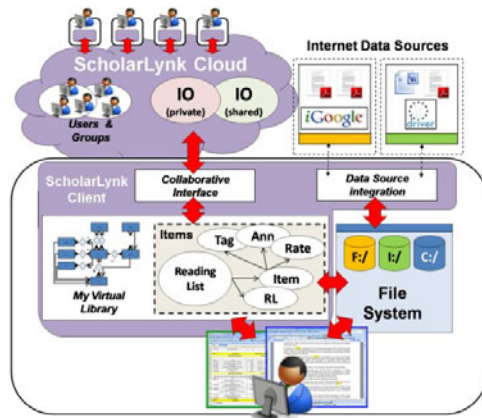


Fig. 3. ScholarLynk architecture

MVL; (ii) *Integration of the MVL environment with* desktop applications such as Microsoft Office and Windows Explorer. The MVL is a client application that abstracts over the file system to enable management of information objects.

For example, while working on a Word file, the user can add the document to a reading list, thus making it an item in the MVL, through the ScholarLynk application toolbar (similar to the desktop toolbar shown in Fig. 1 but showing contextual information about the resource). MVL is also integrated with Windows Explorer in order to deliver MVL item management at the file system level through right click and drag and drop operations. For a given item (e.g., file or directory), users can visualize the reading lists it belongs to and the interpretation associated with the file in the context of the reading lists.

The client offers a range of functions for information object and user management:

- *Information object management.* Based on their access rights, users can manage reading lists and can tag, comment on, and rate information objects.
- *Information object sharing.* Information objects created in the MVL can be shared by the user. *Publication profiles* are defined by the users and specify the publishing options for each object. An information object can be *unpublished* (local) or *published* (sent to the cloud). The publishing properties are:
 - *Private/shared:* sharing can be at the level of users (private), at the level of one or more user groups, or *available to all* (public)
 - *Access rights:* a shared information object can be visualized only, or can be edited with various rights, e.g., add or remove tags, if reading list, add or remove items.
- *User profile management.* Profile management mirrors social network profiles, where user can specify personal information, create and participate in user groups. From the profile, a user handles a wall into which messages can be written manually or automatically, e.g., when the user executes actions on published and shared (at least for viewing) information objects or user profile.

- *Publish and subscribe.* Through the MVL, users may also set a number of *follow* actions (i.e., subscriptions to topics) for the *observable objects* (e.g., information objects, users, groups are observable objects) exposed through the ScholarLynk environment. An observable object collects all *facts* that are directly or indirectly related with it during its life-cycle. For example, the action of adding an item Y to a reading list X affects at least three observable objects and results in the fact: “Item Y has been added by user W to reading list X”. Facts are returned in the form of notification messages to the MVL panel of subscribing users. Other notifications can also be supported, e.g., by email, SMS, Twitter, Facebook, etc.

ScholarLynk Cloud. The ScholarLynk Cloud is a centralized and web accessible environment to which clients connect in order to *publish* their information objects and receive notifications about their follow actions. To this aim the Cloud manages: (i) user and group profiles, (ii) published information objects and their interpretations together with their access rights, and (iii) users’ subscriptions, i.e. follow actions for observable objects. User access to the Cloud may also be possible through a web portal providing all functionalities available from MVL clients.

4.4 ScholarLynk Prototype

We implemented ScholarLynk prototype [16] to test the effectiveness of the proposed researcher-centric approach by measuring the impact of its adoption among a community of researchers. Our intention is to deploy ScholarLynk to the users of the DRIVER Data Infrastructure¹ comprising 2,500,000 metadata records of Open Access publications from over 250 international repositories. To this aim, DRIVER is integrated as a ScholarLynk client web data source.

Implementation of ScholarLynk Cloud. ScholarLynk exploits and extends the D-NET Software Toolkit [15] (web service-oriented infrastructure solution developed as part of the DRIVER-II EC project²) to implement the ScholarLynk cloud. The current prototype makes an assumption that all information objects reside in the cloud and thus no off-line MVL operations are currently enabled. In particular, D-NET is extended with graph database services (based on Neo4J and MongoDB, for payloads) to manage ScholarLynk data model entities. Facebook’s Cassandra will be used to implement efficient storage of facts for observable objects and their delivery as notifications³.

Implementation of ScholarLynk Client. The ScholarLynk client is implemented in Windows Presentation Foundation (WPF) that allows a rich user experience along with the tight platform integration, creating an immersive and collaborative research environment. When the client connects to the cloud (Comet programming⁴) it “boots up” by sending off the user-defined follow actions and downloads information objects and their relative interpretations into its MVL environment. Once on-line, the client supports integration and federated search across data sources, at the moment DRIVER and Google Scholar. The MVL supports users in finding information objects shared

¹ <http://search.driver.research-infrastructures.eu>

² <http://www.driver-community.eu>

³ <http://neo4j.org>, <http://www.mongodb.org>, <http://cassandra.apache.org/>

⁴ http://en.wikipedia.org/wiki/Comet_%28programming%29

by others and managing information objects directly from their desktop environment. Users can view the contents of reading lists and items and set follow actions on them in order to receive notifications of actions regarding them. These are displayed as text messages in the MVL panel.

5 Conclusions

In this paper, we presented an architecture design for supporting scholarly research activities, building on a data model constructed around the concept of a reading list and extended to include heterogeneous resources and their interpretations. The benefits of the ScholarLynk environment include the provision for a unified view and access to desktop and web content, leveraging collective knowledge through exposed social data, and supporting collaboration through reading list sharing and in-context communication. Unlike systems such as Mendeley, ScholarLynk is not application-bound but provides an integrated desktop-based view, facilitating seamless management of resources through tagging, regardless of the service, application, or their storage location. We believe that these are key requirements for making digital libraries more accessible, more personal, and socially aware. Our future plans include the evaluation of our ScholarLynk prototype through deployment in user studies. Collected usage data, user feedback, and findings from the user studies will guide future design and development.

References

1. Alford, M.L., Mendes, E.: Scholarly research process: investigating the effects of link type and directionality. In: Proc. of HYPERTEXT 2009, pp. 99–108. ACM, New York (2009)
2. Bishop, A.: Digital libraries and knowledge disaggregation: The use of journal article components. In: Proc. of ACM Conference on Digital Libraries, pp. 29–39. ACM Press, New York (1998)
3. Borgman, C.L., Furner, J.: Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology* 36 (2002)
4. Dang, D.-T., Tan, Y.F., Kan, M.-Y.: Towards a Webpage-Based Bibliographic Manager. In: Buchanan, G., Masoodian, M., Cunningham, S.J. (eds.) ICADL 2008. LNCS, vol. 5362, pp. 313–316. Springer, Heidelberg (2008)
5. Ellis, D.: Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly* 63(4), 469–486 (1993)
6. Farooq, U., Ganoë, C.H., Carroll, J.M., Giles, C.L.: Supporting distributed scientific collaboration: Implications for designing the CiteSeer collaboratory. In: Proc. of 40th Annual Hawaii International Conference on System Sciences. IEEE Comp. Society, Los Alamitos (2007)
7. Henning, V., Reichelt, J.: Mendeley - A Last.fm For Research? In: Proc. of the 2008 Fourth IEEE International Conference on Escience, pp. 327–328. IEEE Comp. Society, Los Alamitos (2008)
8. Hull, D., Pettifer, S.R., Kell, D.B.: Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web. *PLoS Comput. Biol.* 4(10) (2008)
9. Kolb, D.: Association and argument: Hypertext in and around the writing process. *NRHM* 11, 7–26 (2005)

10. Marlow, C., Naaman, M., Boyd, D., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proc. of HYPERTEXT 2006, pp. 31–40. ACM, New York (2006)
11. Nakakoji, K., Yamamoto, Y., Akaishi, M., Hori, K.: Interaction design for scholarly writing: hypertext representations as a means for creative knowledge work. *New Rev. Hypermedia Multimedia* 11(1), 39–67 (2005)
12. Oleksik, G., Wilson, M.L., Tashman, C., Mendes Rodrigues, E., Kazai, G., Smyth, G., Milic-Frayling, N., Jones, R.: Lightweight tagging expands information and activity management practices. In: Proc. of CHI 2009, pp. 279–288. ACM, New York (2009)
13. Palmer, C.L., Tefteau, L.C., Pirmann, C.M.: Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development. In: Report commissioned by OCLC Research (2009), <http://www.oclc.org/programs/publications/reports/2009-02.pdf>
14. Volda, S., Mynatt, E.D., Edwards, W.K.: Re-framing the desktop interface around the activities of knowledge work. In: Proc. of UIST 2008, pp. 211–220. ACM, New York (2008)
15. D-NET Project, <http://www.d-net.research-infrastructures.eu>
16. ScholarLynk Project, <http://research.microsoft.com/scholarlynk>

CritSpace: A Workspace for Critical Engagement within Cultural Heritage Digital Libraries

Neal Audenaert, George Lucchese, and Richard Furuta

TEES Center for the Study of Digital Libraries
Texas A&M University
College Station, TX, 77840
{neal, eyce900, furuta}@csdl.tamu.edu

Abstract. Cultural heritage digital libraries hold promise both as a new tool for representing the complex information structures frequently found in the humanities and social sciences and as interactive environments that enable scholars to work with this information in new ways throughout the research project. Much attention has been paid to digitization, textual encoding, metadata and dissemination of digital cultural heritage data. Scholars now routinely turn toward electronic sources as a first step in their information finding process. Considerably less attention, however, has been devoted to understanding how to support the formative stages of scholarly research.

In this paper, we highlight our finding from a formative user study of scholarly analysis of source documents in several different fields. We discuss the implications of these results for our current research into designing a web-based creativity support environment for cultural heritage digital libraries.

1 Introduction

In this paper we describe our ongoing work to design a creativity support environment (CSE) for humanities scholars in the formative stages of their work with primary source documents. The materials of humanities scholarship are complex, intricately interconnected, and, often, visual in nature. Printed presentations of this material have evolved over centuries that account for some of these challenges, but the resulting books are notoriously difficult to use and expensive to produce [13]. Consequently, cultural heritage digital libraries are a powerful tool for dealing with the inherent complexities of these materials.

The specific advantages of digital libraries to many areas of practice within the humanities are now well-documented. These include both increased expressive power for representing complex networks of relationships and the use of computational tools and interactive environments to help researchers ask new questions [8].

Despite the rapid growth in technology and its continuing impact on the ways cultural heritage materials are presented and accessed, many aspects humanities work practices remain unchanged. Digital libraries, with varying degrees of success, have achieved the goal of supporting tasks Unsworth calls “scholarly primitives:” discovering, annotating, comparing, referring, sampling, illustrating, and representing [16].

These tasks are the building blocks scholars use to analyze, reflect, interpret, and understand the material representations of the communities and cultures they study.

These higher-order activities have proven more challenging to support. In describing when scholars should not use the TEI, Lavagnino notes, “the appropriate scholarly tools for the early exploratory stages of a project may be pen and paper, or chalk and a large blackboard, or a word processor” [6]. Cultural heritage digital libraries are seen as a means for publishing and disseminating “completed” work.

Within the digital humanities community, research into digital libraries has emphasized defining the metadata formats and encoding standards needed to prepare and disseminate the results of scholarly research—frequently digital editions or comprehensive archives. Digital libraries and digital representations of texts are seen as a publishing model that overcomes some of the limitations of the book as a machine for knowledge. Consequently, work in this area has focused on the adequacy of these models for representing scholarly knowledge rather than on the interfaces used to present that knowledge or the needs of readers.

In this paper, we argue that digital libraries serve (or should serve) as both a new venue for publishing and finding material and as environments that support the formative stages scholarly research. In fulfilling this latter role, we envision user interfaces that support not only information consumers but also that provide vital support to the scholars responsible for iteratively developing the content to be contained in the library. In the following pages, we summarize the results of a series of semi-structured interviews that we conducted in order to identify common work-practices between scholars who work with primary source documents from many different disciplinary perspectives. We then describe the implications of the findings and prior work in designing CSE’s to the task of develop a web-based application to support formative research involving cultural heritage materials.

2 Understanding Humanities Scholars

To maximize the usefulness of the large, heterogeneous collections of digital resources, scholars need interactive environments that support the formative stages of their research as well as the terminal stages. But what do these early stages of research look like? Unfortunately, little is currently known about the work practices of scholars from an external perspective [2]. What is known has largely been motivated from the library sciences communities needs to better support information seeking behaviors [5][12]. More recent work has focused the impact of technology and the design needs to support effective search and browsing [3][14][15]. This work offers useful insights into the design and implantation of these resources, but continues to emphasize digital libraries as a repository for ‘finished’ works of scholarship leaving the critical tasks of creating those works unaddressed.

In our efforts to design tools that support scholars’ formative research practices, we conducted our own investigation their use of source documents—texts for which they need to see the original (or a high-fidelity facsimile of it) rather than a transcribed form. Our objective in this study was to understand why scholars uses these documents (for example, is it merely an emotional attachment to traditional methods without any scholarly merit or do they have specific research needs that require access

to source material) and the types of tasks they are engaged in. Instead of conducting an in-depth investigation of specific work practices, we chose pursue a high-level perspective that would inform the design of general purpose tools. These tools can be used across multiple domains and customized as needed to support specific needs. We report the details of our results elsewhere [1] and summarize them here for context.

Scholars use the visual information in these documents both to gain a high-level perspective of a particular work and as a source for detailed investigations, often on seemingly esoteric matters such as how worn the type was when it was used to print a particular copy of a book. They use insights gained from this study to understand not only the work itself (for example, to reconstruct a lost original text from multiple conflicting witnesses) but also the context in which the object was produced. Scholars are interested in a complex ecosystem of ideas. To make sense of this ecosystem, they use contextual information to inform their study of physical objects and use those physical objects in turn to inform their understanding to the embedding social, political, intellectual, and economic context.

Notably, even when scholars professed little interest in source material other than as the only available representation of some bit of information, they internalized and frequently referenced visual features of these materials when discussing them. While they may profess that their readers would have no use for this information, it clearly plays an important, though implicit, role in the development of their own ideas.

Unsurprisingly, we found that scholars read widely, scouring all available material in order to deeply internalize the body of knowledge surrounding specific research themes. This internalized knowledge forms the basis from which they begin to articulate their own interpretive voice. As part of this process, note taking plays a key role. Notes are used to record key observations and facts and to provide an external record of the scholar's own ideas and conjectures. Frequently, these notes are conceived of as an integral part of the writing process. They will become the footnotes, paragraphs and sections of a book or journal article. Consequently, scholars choose record and organize notes based on their formative ideas about the structure and content of a writing project that they have in mind.

Stemming in part from their need to quickly integrate notes with the final form of a writing project, many scholars strongly prefer to record their notes electronically. This observation is particularly relevant for our own goals. Scholars do not content themselves with 'pencil and paper' as Lavagnino suggested. Instead, they use technology on an ongoing basis beginning with the earliest stages of their research. The technology they use, however, typically consists of basic word processing tools. While these tools may be adequate for some tasks, they offer little support tailored to their specific needs and no integration with existing digital resources.

3 CritSpace: A Creativity Support Environment for Scholars

In this section we provide an overview of the interaction paradigm we have adopted in designing CritSpace and then discuss how that basic design supports key requirements identified in our user study. We are currently in the process of re-architecting our initial prototype and discuss how some of the features of this new version reflect and support the needs we have identified in our user study.

3.1 Overview

Prior research into creativity support environments [9][11] demonstrates that the early stages of scholarship require tools that facilitate the expression of ambiguous, partial, and emerging knowledge structures—allowing scholars to discover and articulate their own perspectives through exploratory interaction with copious amounts of data. To support this basic requirement, we began by designing CritSpace as a two and half dimensional work-space for scholars to use to collect and organize content from a digital library and to author new content (primarily notes and user-supplied metadata). Objects can be freely positioned within this workspace and visual properties such as background color, border style, font and opacity can be manipulated in order to provide added dimensions to visually organize the information space. The workspace can be arbitrarily large and panned to work with areas that are too large to be displayed on a single screen. The workspace is embedded within a frame that can be used to display information and interfaces that should be always present such as a general purpose writing area, metadata about the currently selected content or menu items.

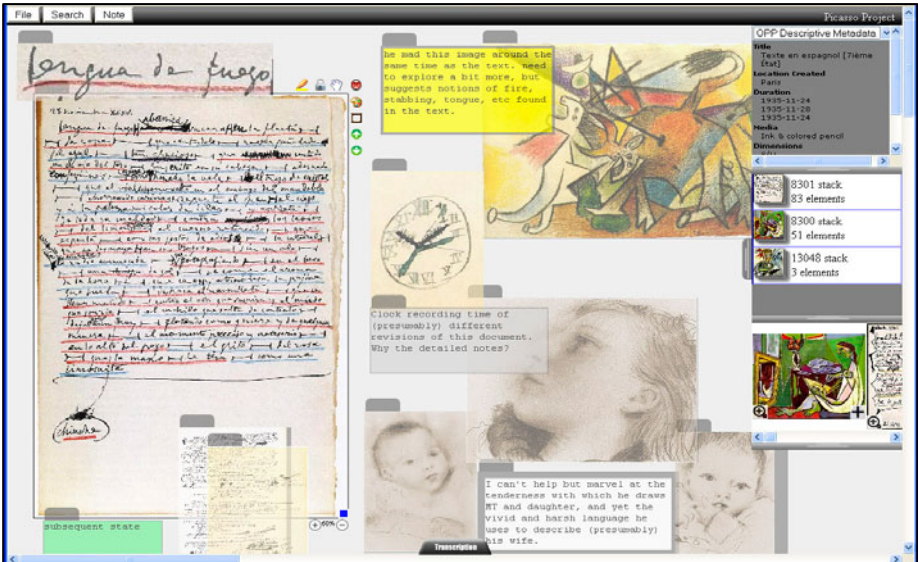


Fig. 1. Workspace Overview

Since the web has become the de-facto standard for delivering digital library content, the CritSpace front-end is implemented as a rich Internet application using JavaScript, HTML and CSS. The front-end is designed to interact with heterogeneous server-side data sources via AJAX. Figure 1 shows an early prototype of the system with images and documents taken from the Online Picasso Project. We are currently working on a second-generation prototype. Our primary goal for the core system is to build an interactive environment that provides a lightweight, flexible interface for gathering and organizing information. We want to implement an interaction paradigm

that allows users to record thoughts quickly as they reflect on and work with information in the environment. Prior research on similar systems has demonstrated that spatial models for knowledge organization meet these objectives well [10].

One innovation of the CritSpace architecture is the separation of the basic objects of manipulation (we call them panels) from the specifics of the content to be displayed. In the prototype shown in figure 1, the panels can be easily identified by the tab controls in their upper left hand corners. The selected panel is shown with small control icons arranged around its upper right hand corner. These panels function as mini-applications that represent an arbitrarily wide range of “content objects” including basic objects such as images and notes (as shown in figure 1) or more complexly structured objects like facsimile documents or note cards with multiple fields.

Panels may fire events (and listen to the events fired by other panels) to enable interaction between the panels used to display different content objects. For example, to facilitate the comparison of similar works, we plan to implement linkable ‘facsimile’ panels. Turning the page in one book will fire a ‘pageTurn’ event that the linked the linked panel can listen for and update its display appropriately.

By designing the architecture for event driven communication and to separate visual manipulation from rich interaction with different content types, we are able to quickly to implement and test different interaction strategies within the basic framework of visually-based interactive environment.

One of CritSpace’s most obvious, and most important, features is the ability to allow scholars to interact visually with visual source material. By giving primacy to image-based content and spatial/visual metaphors for organizing information, we have designed CritSpace to help scholars explore and reflect on the visual features of the objects they study. We are currently working to integrate the lessons we learned from our user-study by designing and implementing three content display panels. These panels are designed to take advantage of the distinctive user interaction style afforded by the CritSpace system and to embody and evaluate tools to address specific findings from our study.

3.2 Tzivi: Tiled Zoomable Image Viewer

One of the most basic types of information content in the CritSpace environment is images. Images serve as the raw material for developing a holistic impression of a document and will be subjected to detailed examination. When transcriptions and other derived sources aren’t yet available, page scans or digital photographs will provide the first digital form of a document.

Interaction with image-based content requires more than simply manipulating thumbnail images that float in a two dimensional workspace. In designing an improved image panel, we want to allow scholars to see both low-resolution overviews of an entire page (or other type of object) and to zoom in to examine details in high-resolution images. The overview display mode is intended to encourage development of holistic impressions. It also provides a low-fidelity representation that serves as a visual anchor to promote reflection and remind users of possible relationships between multiple different documents. The details-oriented perspective allows users to focus on individual regions of interest and examine specific features in detail.

To achieve this we have implemented a tile-based zoomable image viewer, Tzivi, that is conceptually similar to Zoomify or Google Maps. This is now a reasonably common approach for viewing high-resolution images. Most implementations, however, emphasize the inspection of a single image and provide users with a fixed viewport that dominates the screen real estate. In contrast, integrating our Tzivi application into the CritSpace framework will allow users to freely resize and move the image viewport as well as zoom in/out on the image. This will allow users to simultaneously work with many different high-resolution images in a single workspace.

Another key feature of the Tzivi panel is its ability to allow the user to add multiple layers of annotations to the image. These annotations may define rectangular regions, polygons or ellipses. Alternatively, they may be point of interest markers or notes. Annotations may be used to specify structured metadata about a particular region of an image or simply serve as a visual marker. This provides scholars with flexible notation support to record observations tied to specific features of an object.

3.3 Facsimile Viewer

Building on our base image-viewer, we next turn our attention to supporting image-based facsimile editions. Most documents of scholarly interest consist of more than one page and this additional internal structure requires support. Accordingly, we are designing a facsimile panel that supports digital facsimiles and displays pages either as a series of thumbnails (to promote rapid browsing in cases where a transcription is unavailable) or as individual pages using a modified form of the Tzivi panel that supports page-based navigation. Users will be able to seamlessly switch between these two navigational modes. By allowing facsimiles of related documents to be linked, this panel will facilitate the side-by-side comparison of related documents, a critical task in understanding the textual transmission of documents.

3.4 Faceted Browsing

Instead of implementing search and browsing features as part of the core CritSpace framework, we are designing these tools as panels. This allows users to open several search panels to create different sets of search results. These results are then persisted in the workspace just like any other content object and may be referenced or revised at any time. Unlike traditional search tools, these search panels define a set of results to which content objects may be added or removed manually. Like the facsimile viewer, objects in the collection may be displayed either as a sequence of thumbnails or as individual images that can be navigated sequentially.

Faceted browsing has been a mainstay of exploratory search interfaces and research in cultural heritage digital libraries has confirmed its usefulness in this environment [18]. Specifically within the context of the findings from our study, we see faceted browsing as a key tool to help scholars conduct comprehensive surveys of available material. By adding support for manually refining search results and for persisting the results of multiple searches within a single workspace, we expect to further improve on this technique.

4 Summary

Cultural heritage digital libraries hold tremendous promise to supplement physical libraries as the laboratory of humanities. These libraries enhance our ability to store and find information, support automated tools that assist in analyzing, summarizing and visualizing that information, and create infrastructure to support collaborative and cooperative work practices. These advances are critical but insufficient. While we must continue to invest in publication of the finished results of humanities scholarship, we must also consider how to support the analytical and creative processes that give birth to those results.

We have proposed the spatial information management strategies embodied by CritSpace as one direction for supporting the early stages of humanities scholarship along with some supporting technologies. We believe these tools will support scholarly use of the content contained in cultural heritage digital libraries, in particular, enhancing their ability to rapidly pose, see and reflect on new ideas and hypotheses. We are currently re-engineering our initial prototype interface and expect to begin formal user studies in the fall of 2010. These studies will consist of medium to long-term use of the CritSpace framework within the context of affiliated research groups. Based on the results of these studies, we will use the framework as the basis for ongoing research on how to best support specific needs of different research communities within the humanities and other disciplines.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0534314.

References

- [1] Audenaert, N., Furuta, R.: What humanists want: how scholars use primary source documents. In: Joint Conference on Digital Libraries, JCDL 2010 (2010) (forthcoming)
- [2] Borgman, C.L.: The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly* 3(4) (2009)
- [3] Buchanan, G., et al.: Information seeking by humanities scholars. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 218–229. Springer, Heidelberg (2005)
- [4] Csikszentmihalyi, M.: *Creativity: Flow and the Psychology of Discovery and Invention*. HarperCollins, New York (1996)
- [5] Ellis, D.: A behavioural model for information retrieval system design. *Journal of Information Science* 15(4/5), 237–247 (1989)
- [6] Lavagnino, J.: When not to use TEI. *Electronic Textual Editing*. Modern Language Association, New York, http://www.tei-c.org/About/Archive_new/ETE/Preview/lavagnino.xml (viewed May 2010)
- [7] Marshall, C.C., Shipman, F.M.: Spatial hypertext and the practice of information triage. In: *ACM Conference on Hypertext (Hypertext 1997)*, pp. 124–133 (1997)

- [8] McGann, J.: The Rational of Hypertext. In: Sutherland, K. (ed.) *Investigations in Method and Theory*, pp. 19–46. Clarendon Press, Oxford (1997)
- [9] Shneiderman, B.: Creativity support tools: accelerated discovery and innovation. *Communications of the ACM* 50(12), 20–32 (2007)
- [10] Shipman, F., Marshall, C.: Spatial hypertext: an alternative to navigational and semantic links. *ACM Computing Surveys* 31(4es) (1999)
- [11] Shipman, F.M., Marshall, C.C.: Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Computer Supported Cooperative Work* 8(4), 333–352 (1999)
- [12] Stone, S.: Humanities scholars: information needs and uses. *Journal of Documentation* 38(4), 673–691 (1982)
- [13] Tanselle, G.: Critical Editions, Hypertexts, and Genetic Criticism. *Romantic Review* 86, 581–593 (1995)
- [14] Tibbo, H.R.: Primarily history: historians and the search for primary source materials. In: *Joint Conference on Digital Libraries (JCDL 2002)*, pp. 1–10 (2002)
- [15] Toms, E.G., O’Brien, H.L.: Understanding the information and communication technology needs of the e-humanist. *Journal of Documentation* 64(1), 102–130 (2008)
- [16] Unsworth, J.: Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? In: *Paper Given at Humanities Computing: Formal Methods, Experimental Practice*, King’s College, London (2000), <http://www3.isrl.illinois.edu/~unsworth//Kings.5-00/primitives.html> (viewed May 2010)
- [17] Yamamoto, Y., Nakakoji, K., Aoki, A.: Spatial Hypertext for linear-information authoring: Interaction design and system development based on the ART Design principle. In: *ACM Conference on Hypertext and Hypermedia (Hypertext 2002)*, pp. 35–44 (2002)
- [18] Zhang, J., Marchionini, G.: Evaluation and evolution of a browse and search interface: Relation Browser++. In: *2005 National Conference on Digital Government Research*, pp. 179–188 (2005)

German Encyclopedia Alignment Based on Information Retrieval Techniques

Roman Kern¹ and Michael Granitzer^{1,2}

¹ Know-Center, Graz

² Graz University of Technology
{rkern,mgrani}@know-center.at

Abstract. Collaboratively created online encyclopedias have become increasingly popular. Especially in terms of completeness they have begun to surpass their printed counterparts. Two German publishers of traditional encyclopedias have reacted to this challenge and decided to merge their corpora to create a single more complete encyclopedia. The crucial step in this merge process is the alignment of articles. We have developed a system to identify corresponding entries from different encyclopedic corpora. The base of our system is the alignment algorithm which incorporates various techniques developed in the field of information retrieval. We have evaluated the system on four real-world encyclopedias with a ground truth provided by domain experts. A combination of weighting and ranking techniques has been found to deliver a satisfying performance.

1 Introduction

Printed encyclopedias have been the prime source of information for a long time. They are created by experts in their fields and therefore provide a high credibility. Due to their tradition as printed media, encyclopedias follow a particular structure and outline. Space is at prime and therefore articles tend to be terse. Still the articles should contain all available information resulting in a writing style specific to such corpora. Dealing with this kind of language poses an additional challenge for natural language processing (NLP), machine learning (ML) or information retrieval (IR) techniques.

The rise of the Internet and more specifically the popularity of online encyclopedias has put pressure on the producers of traditional printed encyclopedias. While initially there has been doubts whether the new form of collaboratively created resources can match the quality of the established encyclopedias (see for example [1]), more recently traditional publishers have changed their strategy. They have started to put their resources online and also started in parts to allow non-experts to contribute information.

Another way to improve the quality and especially the completeness of an encyclopedic resource is the combination of multiple sources. Starting with two encyclopedias one can create a merged resource that contains the combined and as a consequence a more complete information. The most important step of this

operation is the alignment of articles. Articles about the same person, entity or concept in both encyclopedias should be automatically assigned to each other. Additionally, those articles that only exist in one of the two encyclopedias should be identified and thus create a new entry in the merged corpus.

State-of-the-art methods in NLP and related techniques has not yet reached the level that such an alignment can be conducted completely automatically. Manual intervention of human experts is still necessary in many cases. Prior to developing an encyclopedia alignment system we have set-up a number of goals to achieve:

- The accuracy of the automatic article alignment should be maximized.
- The coverage of automatically aligned articles should be as high as possible, to minimize the number of articles required for manual assignment.
- “Keep the human in the loop” and support the manual alignment by providing an intuitive search infrastructure and useful recommendations.

Our system should also be used in an interactive manner to support manual alignment by domain experts. Therefore the alignment algorithm not only provides a high accuracy, it should also be fast and efficient, as finally that the algorithm should be integrated into a software tool targeted at desktop computers (Figure 1 depicts a prototype of the application). We have decided to choose techniques from the field of information retrieval as the base of our alignment algorithm for a couple of reasons. Search and indexing tools have been developed for a long time and have now reached a mature level. Retrieval algorithms are well studied and their behavior is well understood. In contrast to these methods, the results produced by many supervised classification methods are hardly traceable.

2 Related Work

The most striking characteristic of many articles within traditional encyclopedias is their length. Because of space limitations the majority of all articles are relatively short compared to the covered information.

In [2] an overview of similarity methods for various short contexts is given. Using the categorization presented by the authors, a single encyclopedia article can be classified as head-less context and the alignment can be seen as pairwise comparison to reference samples. To calculate the similarity between two short contexts according to the paper the words can either be directly used or replaced by an representation. The first method is referred as first-order similarity, whereas the second method is called second-order similarity. For the second-order representation the individual word within the context are usually expanded by exploiting an external resource, for example WordNet.

One of the approaches to integrate semantic information via WordNet is presented in [3]. They propose an algorithm to calculate the similarity between individual sentences. The distance between entries within the WordNet graph are taken as proxy for the semantic relatedness of words. Additionally, their algorithm deviates from the unordered bag of words approach by incorporating the word order into their similarity calculation.

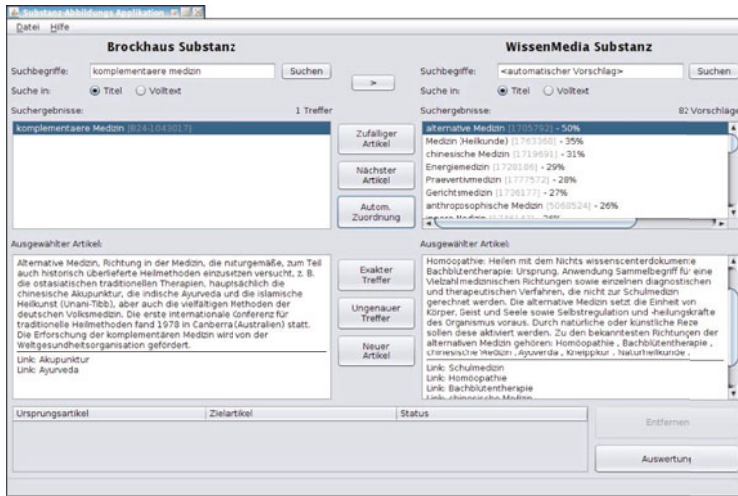


Fig. 1. Prototype of the application that should support domain experts in the process of encyclopedia alignment. The quality and the number of automatically aligned articles has a high impact on the productivity of the experts.

A similar approach is taken by [4] where position information and lexical distance serve as base for the similarity. The performance of this algorithm is compared against a system which employs Latent Semantic Analysis in [5]. In this comparative study they created a benchmark dataset of 30 sentence pairs. At first humans assigned a similarity for each of the sentence pairs which served as ground truth. Finally, they compared the mean similarity from the human judgments with the results of the two approaches. They found that the LSA based algorithm produces a higher correlation than the similarities calculated as described in [4].

Various degrees of similarities are studied in [6], from a broad topical similarity at one end of the spectrum to document identity at the other end. They present various measures to calculate the similarity of sentences and documents, for example the overlap of common words and a *TFIDF* based weighting of shared words. Probabilistic translation models are also investigated in their study together with the DECO system (see [7]) for document similarity. In their evaluation they report the performance of the different similarity measures for various degrees of similarity. For encyclopedia alignment the results at the “same facts” category are the most relevant. For this category the machine translation model and the simple overlap measure provides the best performance.

Beside incorporating resources like WordNet and other thesauri into the similarity calculation, the web has become increasingly popular as knowledge base in recent works. In [8] the authors incorporate the results of web searches into a similarity kernel function. Their method is targeted at finding similar short text snippets, especially substitution candidates for search queries. This approach is further improved in [9] by changing the weighting function and by integrating

machine learning algorithms. Out of the surface matching similarity measures the Jaccard Coefficient fared better than the Overlap and the Cosine similarity in their evaluation.

3 Encyclopedic Corpora

We have had the opportunity to have four encyclopedic corpora at our disposal when developing and evaluating our encyclopedia alignment system. Three of them are from *Brockhaus* and vary in the number of articles and average length of the articles. The fourth corpus, the *WissenMedia* encyclopedia, is comparable in number of articles with the largest of the Brockhaus corpora. Table 1 gives an overview of the statistics of the four datasets.

The task of our alignment system is to merge all articles from the three Brockhaus corpora and the WissenMedia corpus to create one single and complete encyclopedia. For example, the article on the right side of figure 2 should be assigned to the article depicted on the left.

3.1 Anatomy of an Encyclopedia Article

Each article in an encyclopedia consists of multiple parts, whereas the title and the textual content are the most important ones. The main content does not only contain the plain text of the article but also links to other articles and may also feature references to pictures and other media. Depending on the actual encyclopedia it may contain additional annotations not visible in the plain text, for example the number of inhabitants in an article that describes a specific city or country. Other data can be extracted directly from the text, for example the date of birth of a person.

Additionally to the title each article may also contain a sub-title. For articles that represents a person the sub-title contains the person’s first name. The sub-title is sometimes also used for disambiguational purpose, for example the articles with the title “Mexico” also carry the sub-title “city” or “country”. Unfortunately this disambiguation information is not standardized and is used differently in each encyclopedia.

Finally, the article may also carry a wide array of additional meta-data, which is not exploited by our system, for example the pronunciation of the article’s title, assignments to classification taxonomies and hints how the article should look like in printed form.

Table 1. Overview of the statistics of the three *Brockhaus* and the *WissenMedia* encyclopedias. The average length of an article is less then 100 words for each corpus.

	Brockhaus 1	Brockhaus 2	Brockhaus 3	WissenMedia
Number of articles	42,450	94,854	176,963	176,011
Number of unique words	137,929	395,685	840,577	370,076
Number of words	979,958	5,256,761	16,920,079	6,761,156
Average article length	23.16	55.47	95.62	38.42



Fig. 2. The same article in two different encyclopedias. Although the title of the two articles differ, they cover the same topic and should therefore be assigned to each other.

4 Algorithm

The alignment algorithm operates in two stages: a retrieval and a ranking stage. In the first stage for a particular source article a list of candidate target articles is generated. Each of the candidate articles are individually weighted in the second stage. The output of the final stage is a ranked list of possible target articles, where each article's weight ranges between 0 and 1. The highest ranked article is marked as the alignment match for the source article if the weight exceeds a predefined threshold. By choosing a low threshold the number of automatically aligned articles will rise. A high threshold will lead to fewer aligned articles but the number of misalignments will also decline. In the evaluation section we study the influence of this parameter on the systems performance.

4.1 Text Processing

In contrast to the English language, in German noun word-compounds are frequently used. For example the English phrase “coffee maker” can be translated as the single German word “Kaffeemaschine”. In encyclopedias these compound words are even more common than in general German due to the terse nature of articles.

In our system we have implemented two different strategies to deal with these compound words. The first is a simple character n-gram approach that splits words into n-grams of up to 3 consecutive characters. For example the 3-grams of “Kaffeemaschine” are: `kaf aff ffe fee eem ema mas asc sch chi hin ine`

The second approach is more sophisticated. Each tokenized word is first split into syllables based on hyphenation patterns. Each syllable is looked up in a dictionary to detect whether the syllable can be used as a single standalone word. After the syllable has passed this check it is finally stemmed¹. The hyphenation patterns and the dictionary are available from the OpenOffice.org

¹ Stemmer and token splitting algorithms are taken from the open-source Lucene project: <http://lucene.apache.org/java/docs/>

project². The output of this processing for the word “Kaffeemaschine” is: `kaffee
fee kaffeemaschi ma maschi schi`

4.2 Article Facets

The basic data-structure of our alignment algorithm is a search index, which is populated by all articles of the encyclopedias. To capture the different aspects of an article we split the article into different facets:

Title-Exact. The article title is tokenized and normalized. All characters were transformed to lower case, umlauts were replaced with their corresponding digraphs and diacritics were removed. For example the word *Übersee-département* is normalized to: `ueberseedepartement`

Title. The tokenized, normalized title is further processed using one of the two compound words processing algorithms.

Sub-Title. The sub-title (if available) is tokenized and processed like the title.

Content. The body of the article is again split into normalized tokens which were consecutively processed by one of the word-compound processing approaches.

Date. This facet is filled by extracting the birth and death dates out of the content by applying a pattern based approach. This facet is populated only for articles about persons. For example the article about *Johann Wolfgang von Goethe* contains the dates: `*1749 †1832`

Length. This facet is in contrast to the other facets not filled with textual content. It captures the intuition that articles about important concepts tend to contain more words than minor topics. For example the article about famous persons will tend to be longer than articles of people who have not gained huge popularity. Two corresponding articles from two encyclopedias are thus expected to have similar length in relation to the average length of articles within the encyclopedia. The content of the *Length* facet is calculated as defined in equation 1. Important topics have a length ratio close to 1, the ratio for short articles is close to zero and a ratio of 0.5 reflects an article of average length.

$$lengthRatio = \min\left(\frac{length}{2 * averageDocLength}, 1\right) \quad (1)$$

4.3 Candidate Selection and Candidate Weighting and Ranking

Once the search index is created the matching target articles for a source article can be searched. The first step is the selection of a list of possible candidates. Out of the features of the source article a query is build and the top 100 results are selected for further investigation. This query is a disjunction of the facets *Title-Exact*, *Title*, *Sub-Title* and *Content*. In case of the *Content* facet only the 10

² <http://extensions.services.openoffice.org/dictionary>

tokens with the highest weight are taken, using the weighting scheme described in the following section.

In the article weighting step, each candidate is compared with the source article and a similarity score is calculated. The similarity score for each target article is computed by combining the individual similarities of the facets. For each facet - f - out of the set of facets - F - a similarity score is computed for a pair of source and target articles. Not all facets should contribute equally to the final score, thus a predefined boost constant for each facet B_f is incorporated into a weighted mean for the final score:

$$S(s, t) = \frac{1}{B_{sum}} \sum_{f \in F} B_f * boost(score(f, s, t)) * score(f, s, t) \tag{2}$$

The B_{sum} is the sum of all boost constants and serves as a normalization factor for the score to fall between 0 and 1. The $boost()$ function is based on the intuition that similarity scores near the extremes are better suited to assess a similarity or dissimilarity. In the evaluation section a number of boost function are compared against a baseline that just returns a constant value for each similarity score. The actual values for the boost constants B_f were determined on a preliminary test of 100 randomly drawn articles: $B_{TitleExact} = 20, B_{Title} = 25, B_{SubTitle} = 40, B_{Content} = 75, B_{Date} = 50, B_{Length} = 2$.

The most important part of equation 2 is the $score()$ function that calculates the similarity of corresponding facets of two articles. Each facet is transformed into a weighted vector so that different similarity measures can be used, namely: Cosine, City-Block, Euclidean, Jaccard, Dice and Overlap. Distance measures were transformed to similarities via: $sim = 1/(1 - distance)$

To create the weighted term vector for each facet we have integrated a number of weighting functions. The first is a simple *TFIDF* weighting scheme based on the number of articles - N - within the encyclopedia and the number of articles the term t occurs in - $docFreq_t$:

$$weight_{TFIDF}(t) = \log\left(\frac{N}{docFreq_t+1} + 1\right) * \sqrt{termFreq_t} \tag{3}$$

The next term weighting function has been developed using an axiomatic approach to information retrieval, see [10]. This weighting scheme also incorporates the actual length of the article (in this case the number of terms within a facet) and the average length of articles. For the parameter α we used the value 0.32 as suggested by the authors of [10].

$$weight_{Axiomatic}(t) = \left(\frac{N}{docFreq_t}\right)^\alpha \frac{termFreq_t}{termFreq_t+0.5+\frac{docLength}{avgDocLength}} \tag{4}$$

The BM25 retrieval function, see [11], has proven to provide state-of-the-art performance in a number of scenarios. We used the recommended default values for the parameters: $k_1 = 2, b = 0.75$

$$weight_{BM25}(t) = \frac{termFreq_t}{termFreq_t+k_1((1-b)+b*\frac{docLength}{averageDocLength})} * \log\frac{N-docFreq_t+0.5}{docFreq_t+0.5} \tag{5}$$

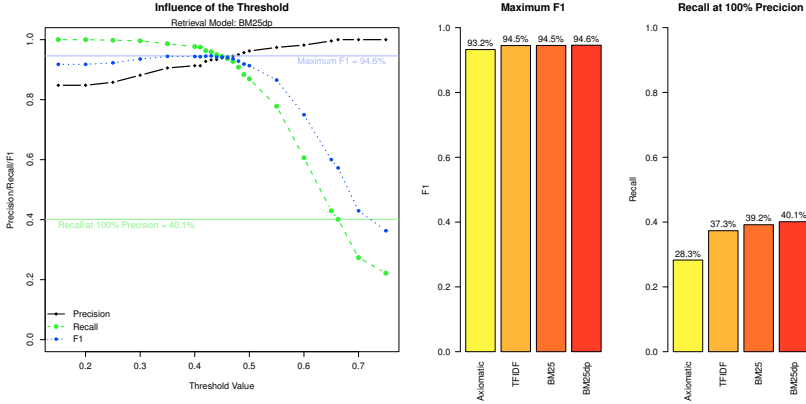


Fig. 3. Precision/Recall/F1 curve for various threshold values for a single retrieval model (BM25dp) on the left side. On the right side: Comparison of all evaluated retrieval models using the two main quality indicators. The modified BM25 retrieval model achieves the highest overall performance.

For the final term weighting function we modified the *BM25* weighting scheme to incorporate the degree of dispersion of terms. The *DP* measure has been proposed by [12] and successfully used by [13] to separate function words from content words. The dispersion degree is low for words with an even frequency distribution, which is expected for words with little semantics but with a grammatical function. The parameter α has been set to -0.3 based on the results of the preliminary tests.

$$weight_{BM25dp}(t) = weight_{BM25}(t) * DP_t^\alpha \quad (6)$$

5 Evaluation and Discussion

The evaluation of our encyclopedia alignment system is based on a ground truth generated by domain experts, which were asked to pick representative articles from their respective domains. The three Brockhaus corpora serve as source and the WissenMedia corpus as target of the alignment. A total of 605 articles were manually processed. For 64 Brockhaus articles the experts have not found a corresponding article in the WissenMedia encyclopedia.

With this ground truth the precision and recall of each configuration of the system can be calculated. The precision is calculated as the number of correctly assigned articles in relation to the total number of automatically assigned articles. Recall defines the ratio of correct assignments to the number of possible assignments (the number of manually assigned articles). The harmonic mean of precision and recall, called F1, is the base of first main indicator for the quality of the results of the algorithm. Running the evaluation with different thresholds

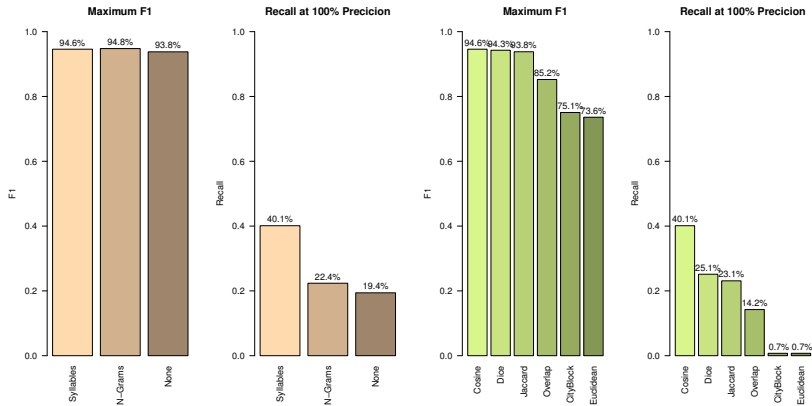


Fig. 4. Comparison of the word-compound processing strategies on the left side, and the performance of the various similarity measures on the right side. The compound splitting method based on hyphenation pattern outperforms the n-gram based method, which is still better than no splitting at all. Out of the similarity measures, the cosine similarity provides by far the best results.

generates a series of F1 measures, see left chart in figure 3. The highest F1 measures defines the best achievable performance when both precision and recall should be equally optimized.

Another characteristic of an evaluation run is the number of aligned articles without a single misalignment. The recall value at the point where the precision reaches 100% captures this property. This measure reflects the usefulness of the configuration if the emphasis lies on optimizing the precision. The higher the recall the less articles have to be manually postprocessed and therefore this indicator plays an important role when choosing a configuration.

The first components of our system to be compared are the different retrieval models, see figure 3. While all four methods appear comparable when using only the F1 based measure as quality criteria, the recall measure reveals that the axiomatic approach falls behind the other retrieval models in terms of performance. The modified BM25 weighting function, which incorporates the dispersion of terms, appears to provide the best results and for this reason this configuration is taken as baseline for all other evaluations.

The next evaluation compares the consequences of the two different word-compound processing methods on the systems performance together with a configuration without any compound-word splitting, see figure 4. While the F1 measure for the n-gram based method is slightly higher, the syllable based approach achieves a higher recall value and thus is better suited for our use case. Still the n-gram based methods is able to perform better than using no splitting at all. This corroborates the need to process compound-words in the German language not only for encyclopedia alignment, but also in other areas, like for example information retrieval.

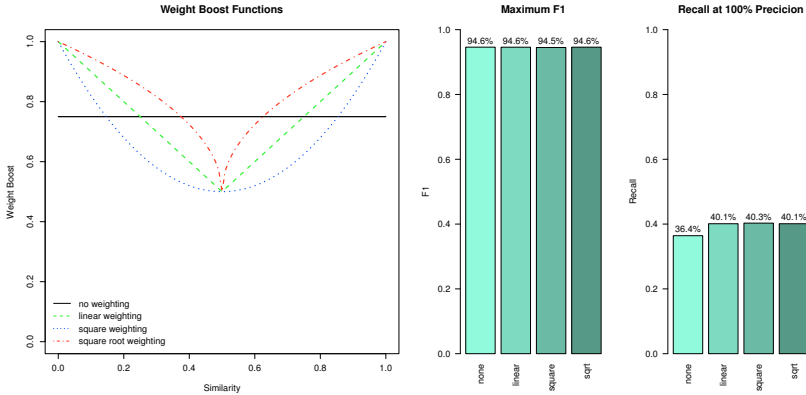


Fig. 5. Weighting functions that capture the intuition that similarities close to the extremes should have a higher impact on the final assignment. The shapes of the boosting function are depicted on the left and the performance indicators on the right. The improved recall at 100% precision indicates that this intuition is indeed sound.

Table 2. Performance indicators for each facet when left out of the similarity calculation. Each facet appears to contribute to the quality of the alignment with the *Content* facet being the most important one.

Facet	Maximum F1	Recall at 100% Precision
Title-Exact	93.4%	40.5%
Title	92.5%	40.9%
Sub-Title	94.3%	39.0%
Content	86.4%	17.7%
Date	94.6%	35.7%
Length	94.4%	39.6%

The results of the evaluation of the different similarity measures are simple to interpret. The cosine measure outperforms all other similarity measures considerably. One reason is the fact that some facets are very sparse, for example the *Title* facet. Applying different similarity measure on different facets could be one possible way to further improve the quality of the alignment algorithm.

Next we tried to assess whether the intuition that similarity measures near the extreme ends are better suited as indicator for similarity. This should especially help in situations where there is an exact match for one of the facets. Figure 5 depicts the baseline (the similarity value has no influence on the boost) and three weight boosting methods. Although the difference between boosting methods appears to be negligible, the boosting approach itself improves the recall by about 4%.

Finally, we investigated the relative influence of each facet. To measure the individual contribution of facets we have repeated the evaluation while removing

one facet at a time. As expected the content of the article is by far the most important factor. Still all other facets contribute to the quality of the result, whereas the two facets generated from the title appear to be slightly redundant. The date and the length information have little influence on the maximum F1 measure, removing them has a pronounced negative effect on the recall at 100% precision measure. Only the combination of all facets provides the best overall performance of our encyclopedia alignment system.

6 Conclusion and Future Work

The automatic alignment of encyclopedic corpora poses a number of challenges. The style of the German language differs from the common language usage because of the terse nature of the articles. Additionally the alignment process should be fast and efficient to be used in an interactive manner. Furthermore, the results produced by the system should be predictable and easy to interpret.

We created such an encyclopedia alignment system by applying techniques developed in the field of information retrieval. Domain experts manually aligned over 600 articles of four real-world encyclopedias. Using this ground truth we evaluated a number of configurations with different retrieval functions, similarity measures and text processing methods. The combination of a modified BM25 weighting function, the cosine similarity and a dictionary based word splitting algorithm provided the best overall performance. This configuration achieved an maximum F1 measure of over 94% and over 40% recall without a single misalignment.

To further improve the quality we could exploit the internal link structure and integrate external resources, for instance thesauri. Incorporating machine translation techniques and language models are among the possible candidates for future improvements.

Although our system has been developed to align articles from different encyclopedias, it should be easy to adapt for other purposes. The detection of duplicates is probably the most obvious application. Other areas are the named entity recognition and disambiguation, which could be integrated into a link recommendation system. Some aspects of our alignment system should not only apply to encyclopedias, but to other textual resources as well. The word-compound splitting method and the dispersion based term weighting should be helpful in other text processing applications as well.

Putting technical aspects aside we believe that our alignment system also serves as good example how science and industry can work together to create solutions and insights beneficial for both sides.

Acknowledgements

We would like to thank Kai-Ingo Neumann and his team at wissenmedia for their support in providing the datasets. The Know-Center is funded within the

Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

1. Rector, L.H.: Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference Services Review* 36(1) (2008)
2. Pedersen, T.: Computational Approaches to Measuring the Similarity of Short Contexts: A Review of Applications and Methods. *CoRR abs/0806.3* (2008)
3. Liu, X., Zhou, Y., Zheng, R.: Measuring semantic similarity within sentences. In: *Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC*, vol. 5, pp. 2558–2562 (2008)
4. Li, Y., McLean, D., Bandar, Z.: Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering* 18(8), 1138–1150 (2006)
5. O’ Shea, J., Bandar, Z., Crockett, K., McLean, D.: A Comparative Study of Two Short Text Semantic Similarity Measures. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) *KES-AMSTA 2008. LNCS (LNAI)*, vol. 4953, pp. 172–181. Springer, Heidelberg (2008)
6. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity Measures for Tracking Information Flow. In: *CIKM 2005: Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 517–524. ACM, New York (2005)
7. Bernstein, Y., Zobel, J.: A Scalable System for Identifying Co-derivative Documents. In: *String Processing and Information Retrieval*, pp. 55–67 (2004)
8. Sahami, M., Heilman, T.: A web-based kernel function for measuring the similarity of short text snippets. In: *WWW 2006: Proceedings of the 15th International Conference on World Wide Web*, pp. 377–386. ACM, New York (2006)
9. Yih, W., Meek, C.: Improving similarity measures for short segments of text. In: *AAAI 2007: Proceedings of the 22nd National Conference on Artificial Intelligence*, pp. 1489–1494. AAAI Press, Menlo Park (2007)
10. Fang, H., Zhai, C.: An exploration of axiomatic approaches to information retrieval. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 480–487. ACM, New York (2005)
11. Robertson, S., Gatford, M.: Okapi at TREC-4. In: *Proceedings of the Fourth Text Retrieval Conference*, pp. 73–97 (1996)
12. Gries, S.: Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), 403–437 (2008)
13. Kern, R., Granitzer, M.: Efficient linear text segmentation based on information retrieval techniques. In: *MEDES 2009: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pp. 167–171. ACM, New York (2009)

Lightweight Parsing of Classifications into Lightweight Ontologies

Aliaksandr Autayeu, Fausto Giunchiglia, and Pierre Andrews

DISI, University of Trento, Italy

Abstract. Understanding metadata written in natural language is a premise to successful automated integration of large scale, language-rich, classifications such as the ones used in digital libraries. We analyze the natural language labels within classification by exploring their syntactic structure, we then show how this structure can be used to detect patterns of language that can be processed by a lightweight parser with an average accuracy of 96.82%. This allows for a deeper understanding of natural language metadata semantics, which we show can improve by almost 18% the accuracy of the automatic translation of classifications into lightweight ontologies required by semantic matching, search and classification algorithms.

1 Introduction

The development of information technologies turned the data drought into a data deluge, which seriously complicates data management and information integration problems. This leads to an increasing importance of metadata as a tool allowing the management of data on a greater scale. The amount of existing attempts to solve the semantic heterogeneity problem shows its importance and reveals the variety of domains where it applies (see [12]). The state of the art algorithms try to solve the problem at the schema or metadata level [3] and their large-scale evaluations [4] show two important directions for improvement: a) increasing the background knowledge [5] and b) improving natural language understanding [6].

Digital library classifications extensively use natural language, both in structured and unstructured form. Natural language metadata (NLM) uses a specific Natural Language (NL), different in its structure from the normal textual domain of language, and the current NL processing (NLP) technologies that are developed for the latter are not well suited for NLM. Thus, they require a domain adaptation to fit the specific constraints of the NLM structure. Moreover, the size of the current datasets [4], ranging from thousands to hundreds of thousands of labels (see Table 1), poses additional requirements on processing speed, as demonstrated by the LCSH and NALT alignment experiment from [7].

In general, the parsing of NLM has applications in many areas, in particular: a) in the *matching* of tree-like structures (such as Digital Libraries classifications or schemas) or lightweight ontologies [8], b) in the *Semantic Classification*

Table 1. Classification datasets’ characteristics

Dataset	Labels	Sample Size	Unique Labels (%)	Levels	Label Length, NL tokens	
					Max	Avg
LCSH	335 704	44 490	100.00	21	24	4.0
NALT	43 038	13 624	100.00	13	8	1.6
DMOZ	494 043	27 975	40.48	12	12	1.8
YAHOO	829 081	132 350	16.70	15	18	2.0
ECL@SS	14 431	3 591	94.51	4	31	4.2
UNSPSC	19 779	5 154	100.00	4	19	3.5

of items of information into hierarchical classifications [9], and in c) *Semantic Search* [10]. All these *motivating applications* require the same steps of natural to formal language translation: a) recognize atomic (language-independent) concepts by mapping NL tokens into senses from a controlled vocabulary, b) disambiguate the senses drawn from the controlled vocabulary and c) build complex concepts out of the atomic ones.

We present the analysis of the NL used in six classifications: **LCSH**¹ (for “Library of Congress Subject Headings”), **NALT**² (for “National Agricultural Library Thesaurus”), **DMoz**³ (for Open Directory Project), **Yahoo! Directory**⁴ (a “catalog of sites created by Yahoo! editors”), **eCl@ss**⁵ (a classification of products and services), **UNSPSC**⁶ (for “United Nations Standard Products and Services Code”), which all illustrate the use of NLM in classifications of information items in different domains. Note that, these datasets contain *subject headings*, *terms* and *category names*, which are all written in NL and which we hereafter refer to as *label(s)*. Table 1 provides some key characteristics of our classifications. We show that the NL used in these datasets is highly structured (see Sections 3 and 4) and can be accurately parsed with lightweight grammars (see Sect. 5). By using parsers based on these grammars, we allow for a deeper understanding of metadata semantics and improve the accuracy of the language to logic translation required by the semantic applications by almost 18% (see Sect. 6) without sacrificing performance.

2 State of the Art

The work available in the semantic web and Digital Libraries is often based on reasoning in a formal language (FL). However, users are accustomed to a NL

¹ <http://www.loc.gov/cds/lcsh.html>

² <http://agclass.nal.usda.gov/>

³ <http://dmoz.org>

⁴ <http://dir.yahoo.com/>

⁵ <http://www.eclass-online.com/>

⁶ <http://www.unspsc.org/>

and it is difficult for them to use a formal one. A number of approaches has been proposed to bridge the gap between formal languages and NL classifications.

Controlled languages (CLs), such as Attempto [11], have been proposed as an interface between NL and first-order logic. This, as well as a number of other proposals based on a CL approach [12,13], require users to learn the rules and the semantics of a subset of English. Moreover, users need to have some basic understanding of the first order logic to provide a meaningful input. The difficulty of writing in a CL can be illustrated by the existence of editors, such as ECOLE [14], aiding the user in CL editing.

CLs are also used as an interface for ontology authoring [13,15,16]. The approach of [15] uses a small static grammar, dynamically extended with the elements of the ontology being edited or queried. Constraining the user even more, the approach of [16] enforces a one-to-one correspondence between the CL and FL. The authors in [13], following a practical experience, tailored their CL to the specific constructs and the errors of their users. Some of these and other CLs have been critiqued [17] due to their domain and genre limitations.

For querying purposes, [18] proposes an NL interface to the ontologies by translating NL into SPARQL queries for a selected ontology. This approach is limited by the extent of the ontology with which the user interacts. Another way to bridge the gap between formal languages and NLS is described in [19], where the authors propose to *manually* annotate web pages, rightfully admitting that their proposal introduces a “chicken and egg” problem. The approach described by [20] for automatically translating hierarchical classifications into OWL ontologies is more interesting, however, by considering the domain of products and services on the examples of eCl@ss and UNSPSC, the authors make some simplifying domain-specific assumptions, which makes it hard to generalise.

Differently from the approaches mentioned above, our work does not impose the requirement of having an ontology, the user is not required to learn a CL syntax, and we do not restrict our considerations to a specific domain. This article develops the theme of [6], improving it in several ways, such as extending the analysis to a wider sample of metadata and introducing a lightweight parser.

3 Part-Of-Speech Tagging

Parts of speech (POS) tags provide a significant amount of information about the language structure. The POS tagging is a fundamental step in language processing tasks such as parsing, clustering or classification. This is why we start our analysis with a look at the POS tags of our classifications.

A random subset of each dataset (see Table 1) is manually tokenized and annotated by an expert with the PennTreeBank part-of-speech tag set [21]. We use the OpenNLP toolkit⁷ to automatically annotate the full datasets. First, using the manually annotated subset of each dataset, we test the performance of the standard OpenNLP tokenization and tagging models, which are trained on the Wall Street Journal and Brown corpus data [22], which both contain long texts,

⁷ <http://opennlp.sourceforge.net/>

Table 2. POS tagger performance, Precision Per Label, %

MODEL	D MOZ	eCL@SS	LCSH	NALT	UNSPSC	YAHOO
D MOZ	93.98	14.12	27.54	75.37	49.69	91.87
eCL@SS	48.80	91.28	28.60	28.73	69.65	62.11
LCSH	81.98	48.79	91.38	81.91	68.14	88.16
NALT	46.97	23.61	28.82	96.42	13.21	34.05
UNSPSC	57.07	45.08	22.76	31.03	92.39	75.46
YAHOO	89.54	15.20	34.84	75.04	45.91	97.91
OPENNLP	<i>49.89</i>	<i>19.02</i>	<i>27.26</i>	<i>40.55</i>	<i>33.20</i>	<i>47.44</i>
ALL-EXCEPT	<i>91.59</i>	<i>58.40</i>	<i>53.25</i>	<i>84.77</i>	<i>76.19</i>	<i>94.77</i>
PATH-CV	96.64	93.34	92.64	96.29	92.72	98.35
COMBINED	99.10	99.69	99.24	99.74	99.40	99.68

mostly from newswire. Second, we train our own tokenization and tagging models and analyse their performance. We use the best performing models for the analysis of the full datasets presented in the next section. In addition, we performed an incremental training to evaluate whether our samples are large enough for the models to stabilize and found that the performances of our models stabilize around 96-98% precision per label on the size of our training samples. This shows that a larger manually annotated sample would not provide important accuracy improvements.

We report the results of our experiments in Table 2 where the columns report the dataset on which the experiments are run and the rows the training model used. As baseline, the “OpenNLP” row reports the performance of the standard OpenNLP tagging model. The “all-except” row reports the performance of the model trained on all datasets except the one it will be tested on to show robustness across datasets and on unseen data. The “path-cv” row reports the performance of the model where the labels appearing higher in the hierarchy were included in the context for training. Finally, the “combined” row reports the performance of the model trained on a combination of datasets. The figures on the diagonal and in the “path-cv” row are obtained by a 10-fold cross-validation. We report in bold the best performances. To indicate the percentage of correctly processed labels we report the precision per label.

We observe that NLM differs from the language used in normal texts. To assess whether NLM could be considered a separate language domain, we did cross-tests and took a closer look at the “all-except” row, comparing it with the “OpenNLP” one. In all the cases the performance is higher by a margin of 25%-47%. At the same time, the differences in model performance on different datasets are smaller than between the models. This performance evaluation confirms the difference between the NL used in metadata and in normal texts and it enables us to select the best applicable model for tagging unknown NLM.

As the major reasons for such differences in performance, we see the lack of context in labels which is not an issue in long texts (see average label length in Table 1), the different capitalization rules between metadata and long texts, and the different use of commas. In addition, the POS tags distribution for labels is different from the one in normal texts as, for example, verbs are almost absent in NLM with, on average, 3.5 verbs (VB) per dataset, ranging from 0.0001% to 0.15% of all tokens of the dataset (see Fig. 1).

4 Language Structure Analysis

The training of the part-of-speech (POS) tagger reported in the previous section enabled the study of the language structure of the classification labels. We analysed the labels' language structure by automatically POS tagging each dataset with the best performing model and found interesting repeating patterns.

For instance, the comma is widely used in LCSH and eCl@ss to structure the labels. LCSH labels are chunks of noun phrases, separated by commas, often in reverse order, such as in the label "Dramatists, Belgian" with the pattern [NNS, JJ] covering 4 437 or 1.32% of all labels. There are also some naturally ordered examples, such as "Orogenic belts, Zambia" with the pattern [JJ NNS, NNP], which can be simplified into two noun phrase (NP) chunks [NP, NP] with independent structures. This pattern accounts for 1 500 or 0.45% of all labels.

We studied some other language characteristics as well, such as label length and POS tag distribution, with which, in addition to the patterns, we can derive grammars to generalize the parsing of the labels and simplify the translation to a formal language (see Sect. 5). This study also allows, by revealing the semantics of different pieces and elements of labels' pattern, to code "semantic actions" attached to the appropriate grammar nodes in our lightweight parser to specialize the translation to the specific language used in the dataset.

Our analysis of the label lengths (see Fig. 1) shows that the majority of labels is one to three tokens long. For example, more than half (50.83%) of all the DMOZ labels contain only one token. Two and three tokens labels represent 17.48% and 27.61%, respectively, while the longer labels only occur in less than 5% of the dataset. In comparison, the LCSH dataset tends to contain longer and more complex labels, with only 8.39% of them containing one token, 20.16% – two tokens and about 10-14% for each of 3-, 4-, 5- and 6-token labels; the remaining 11.45% of labels contain more than 6 tokens. Differently to LCSH, almost all the NALT labels are one and two tokens long. The amount of labels longer than 9 tokens in all datasets is less than 1% and we omit it from the graph.

Fig. 1 shows also the distribution of POS tags. We included all the tags that occur in more than 1% of all the tokens in any of the datasets analysed. Out of the 36 tags from the PennTreeBank's tagset [21], only 28 tags are used in the NLM datasets that we analysed. For comparison, we include POS tag distribution in normal text, represented by the Brown corpus [23].

⁸ POS tags: NNS: plural noun, JJ: adjective, NNP: proper name, CD: cardinal number.

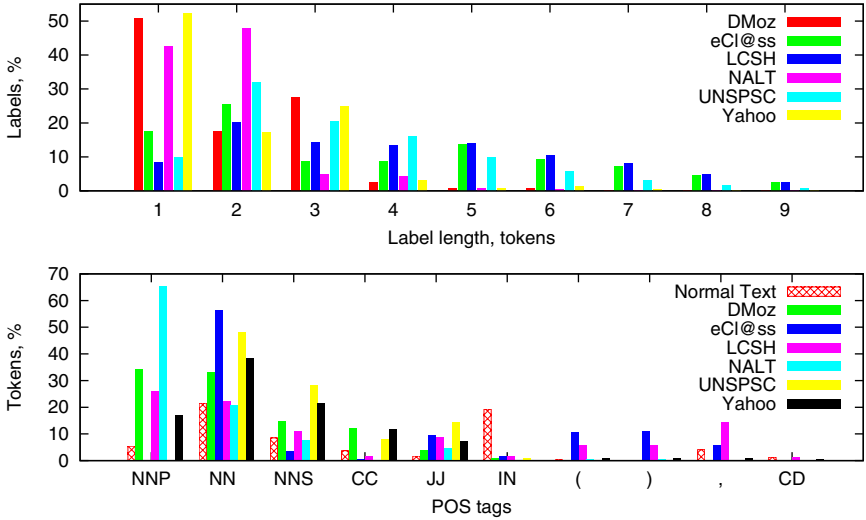


Fig. 1. Distributions of label lengths and POS tags

We observe that all the datasets, except Yahoo, use less than 20 tags in total (see Table 3). Among the top ones are proper nouns (NNP, NNPS) and common nouns (NN, NNS), adjectives (JJ, JJR, JJS), conjunctions (CC), prepositions (IN) and punctuations (“,” and “(”, “)”). A small amount of verbs is present, used as modifiers in the past form (VBD, max 0.0002%) and in the gerund form (VBG, max 0.08%).

Table 3. Metadata language characteristics

Dataset	Tags	Patterns	90% Coverage	Top Pattern
LCSH	20	13 342	1 007	NNP NN
NALT	16	275	10	NNP NNP
DMOZ	18	975	9	NN
YAHOO	25	2 021	15	NN
ECL@SS	20	1 496	360	NN NN
UNSPSC	18	1 356	182	NN NNS

In each dataset we found specific repetitive combinations of POS tags (referred to as patterns). Table 3 shows some characteristics of the language used in classifications with regard to these patterns. The column “90% coverage” shows count of POS tag patterns required to cover at least 90% of the dataset.

A qualitative analysis reveals more details. For example, labels are almost exclusively noun phrases. DMOZ category names are clearly divided into the

“proper” and “common” categories, which was noted in [6]. However, this is not the case for all datasets. Also a noticeable presence of round brackets is explained by their use as a disambiguation and specification tool, as illustrated by the labels “Watchung Mountains (N.J.)” and “aquariums (public)”, which, if treated properly, helps in the formal language translation procedure.

When studying the LCSH patterns at a chunk level (using commas as separators) we can identify 44 groups of chunk-patterns, where many chunks bear clear semantics. For example, the pattern [NNP NNP, NN CC NN, CD] of the label “United States, Politics and government, 1869-1877”, when seen at a chunk level transforms into [geo, NP, time], where “geo” stands for a geographical proper name, “NP” stands for a noun phrase, and “time” stands for a time period.

5 Lightweight Parsing

The parsing of labels in higher level structures can provide a better understanding of their semantics and thus to process them in a more meaningful notation for the computer. Following motivating application a) from Sect. 1, we want to use the S-Match algorithm [24] to align different classifications, such as in the experiment described in [7] and thus need a translation in a lightweight ontology, which. This allows, for example, for the automatic integration of existing heterogeneous classifications.

Rule-based parsers use manually created linguistic rules to encode the syntactic structure of the language. These rules are then applied to the input text to produce parse trees. In long texts parsing, these have been disregarded because of two main disadvantages: they require a lot of manual work to produce linguistic rules and they have difficulties achieving a “broad coverage” and robustness to unseen data. To tackle these problems, state of the art statistical parsers, such as [25], infer grammar from an annotated corpus of text. However, this approach requires a large annotated corpus of text and a complicated process for tuning the model parameters. Moreover, producing a corpus annotated with parse trees is a much more costly and difficult operation than doing a basic annotation, such as POS tagging.

However, as we have seen in the previous section, in NLM, the language used is limited to (a combinations of) noun phrases. Hence, we need a limited coverage, which simplifies the construction of the rules. Therefore we use a simpler approach and manually construct a grammar for parsing. This requires having only an accurate POS tagging and some structural information of the language, which are provided by the analysis we described in the previous sections. We use a basic noun phrase grammar as a starting point for our grammars. Analyzing the POS tag patterns we modify this grammar to include the peculiarities of noun phrases as they are used in NLM, such as the use of commas and round brackets for disambiguation and specification (see examples in Sect. 4).

We have developed a set of lightweight grammars for the datasets discussed in this paper. The grammars we constructed can be divided into two categories: “simple” ones with nine and ten rules (DMoz, eCl@ss and UNSPSC) and a

Table 4. Grammar characteristics

Grammar	Rules	Coverage (%)		Parsing Mistakes (%)	
		Patterns	Labels	POS Tagger	Grammar Rules
LCSH	17	92.96	99.45	49.59	47.94
NALT	15	59.27	99.05	80.35	13.30
DMOZ	9	90.95	99.81	85.98	11.01
YAHOO	15	65.31	99.46	70.90	20.50
ECL@SS	9	67.45	92.70	44.17	47.93
UNSPSC	10	70.58	90.42	25.01	65.70

“complex” ones with 15 and 17 rules (Yahoo, NALT and LCSH). Table 4 provides details about the grammar coverage.

One can note that in all cases we have a high coverage of the dataset labels, more than 90% in all cases and more than 99% in four cases. If we look at the pattern coverage we notice a slightly different picture. For NALT, Yahoo, eCl@ss and UNSPSC, we have only 60% to 70% coverage of the patterns. This can be explained by Table 3 where, for instance, only around 1% of the patterns already cover 90% of the labels in NALT. This shows how a small amount of the labels uses a large variety of language construction while the majority of the NLM uses highly repetitive constructs.

Our analysis shows that the main reason for the lower coverage is a less regular use of language in these four classifications as compared to the other two classifications. We have analysed the mistakes done by the parser and found that they mostly fall into two major categories: POS tagger errors and linguistic rules limitations (see Table 4). This can be explained by the rule-based nature of our parser that makes it particularly sensitive to POS tagger errors. Other parser mistakes are due to the inconsistent (ungrammatical) or unusually complex labels, which could be seen as “outliers”. For example, the “English language, Study and teaching (Elementary), Spanish, [German, etc.] speakers” label from LCSH contains both a disambiguation element “(Elementary)” and a “wildcard” construction “[German, etc.]”.

Fig. 2 shows two examples out of the grammars we produced for the LCSH and UNSPSC datasets. We use Backus-Naur form (BNF) for representing the grammar rules. The LCSH one starts with a top production rule `Heading`, which encodes the fact that LCSH headings are built of chunks of noun phrases, which we call `FwdPhrase`. In turn, a `FwdPhrase` may contain two phrases `DisPhrase` with disambiguation elements as in the example above. The disambiguation element may be a proper noun phrase (`ProperDis`) or a common noun phrase (`NounDis`), surrounded by round brackets. `NounDis` is usually a period of time or a type of object, like “Fictitious character” in “Rumplemayer, Fenton (Fictitious character)” while `ProperDis` is usually a sequence of geographical named entities, like “Philadelphia, Pa.” in “Whitemarsh Hall (Philadelphia, Pa.)”.

<pre> 1 Heading:=FwdPhrase {"", " FwdPhrase} 2 FwdPhrase:=DisPhrase {Conn} DisPhrase 3 DisPhrase:=Phrase {"("ProperDis NounDis")"} 4 Phrase:=[DT] Adjs [Nouns] [Proper] Nouns Foreigns 5 Adjs:=Adj {[CC] Adj} 6 Nouns:=Noun {Noun} 7 Conn:=ConjConn PrepConn 8 Noun:=NN [POS] NNS [POS] Period 9 Adj:=JJ JJR 10 ConjConn:=CC 11 PrepConn:=IN TO 12 Proper:=NNP {NNP} 13 NounDis:=CD Phrase [":" Proper] 14 ProperDis:=ProperSeq ":" Phrase ProperSeq CC ProperSeq 15 Period:=[TO] CD 16 ProperSeq:=Proper [", " Proper] 17 Foreigns:=FW {FW} </pre>	<pre> 1 Label:=Phrase {Conn (Phrase PP\$ Label)} 2 Phrase:=Adjs [Nouns] Nouns 3 Adjs:=Adj {Adj} 4 Nouns:=Noun {Noun} 5 Conn:=ConjConn PrepConn 6 Noun:=NN [POS] NNS [POS] DT RB JJ Proper 7 Adj:=JJ JJR CD VBG 8 ConjConn:=CC , 9 PrepConn:=IN TO 10 Proper:=NNP {NNP} </pre>
--	---

Fig. 2. LCSH (left) and UNSPSC (right) BNF production rules

The core of the grammar is the **Phrase** rule, corresponding to the variations of noun phrases encountered in this dataset. It follows a normal noun phrase sequence of: a determiner followed by adjectives, then by nouns. Alternatively, it could be a noun(s) modified by a proper noun, or a sequence of foreign words.

A comparative analysis of the grammars of different classifications shows that they all share the nine base rules with some minor variations. Compare the rules 4-12 of LCSH with the rules 2-10 of UNSPSC in Fig. 2. These nine rules encode the basic noun phrase. Building on top of that, the grammars encode the differences in syntactic rules used in different classifications for disambiguation and structural purposes. For example, in LCSH, a proper noun in a disambiguation element is often further disambiguated with its type, as “Mountain” in: “Nittany Mountain (Pa. : Mountain)”.

Although very similar to one another, there are a few obstacles that need to be addressed before these grammars can be united into a single one. One of the most difficult of these obstacles is the semantically different use of round brackets: in most cases round brackets are used as a disambiguation tool, as illustrated by the examples mentioned above; however, we also found some examples where round brackets are used as a specification tool, as for instance in the label from eCl@ss: “epoxy resin (transparent)”.

Due to these different semantics, these cases will almost certainly require different processing for a target application. For example, in translating metadata for semantic matching purposes [8], we need to translate the labels of a classification into a Description Logic formula to build up a lightweight ontology. In this application, the disambiguation element “(Pa. : Mountain)” of the label “Nittany Mountain (Pa. : Mountain)” can be used to choose a precise concept “Nittany Mountain” and the element itself is not included in the final formula, while in the specification case of “epoxy resin (transparent)”, the specifier concept “transparent” should be included in the formula in a conjunction with a concept “epoxy resin” that is being specified.

Another obstacle is the different semantics of commas. Sometimes, a comma is used to indicate a sequence of phrases. However, there are cases where the comma separates a modifier in a phrase, written in a “backward” manner, such as illustrated above with a label “Dramatists, Belgian”. In long texts, these differences can be disambiguated by the context, which is almost always missing for NLM.

Despite these differences, our results show that simple and easily customizable grammars can be used to parse accurately most of the patterns found in the state of the art classifications, thus providing extra understanding of the NL without a loss in performance.

6 Evaluation

We have evaluated our approach in a semantic matching application with the dataset from [4] that contains 9 482 labels from a variety of web directories. We have manually annotated all this dataset with tokens, POS tagging information and assigned a correct logical formula to every label. For example, we have annotated the label “Religion and Spirituality” with the POS tags “NN CC NN” and the formula “n#5871157 | n#4566344”, where n#ID point to WordNet synsets for “religion” and “spirituality”, respectively, and | stands for logical disjunction, which was lexically expressed with “and”. The average label length is of 1.76 tokens, with the longest label being of 8 tokens. The most frequent POS tags are singular nouns (NN, 31.03%), plural nouns (NNS, 28.20%), proper nouns (NNP, 21.17%) and adjectives (JJ, 10.08%).

In Fig. 3, we report the accuracy of the translation to description logic formulas, in comparison to the POS tagger performances. We consider the translation to be correct if the resulting formula is logically equivalent to the formula in the manual annotation. We report two different POS tagging models (see Sect. 3): **No Context** that corresponds to the best combined model, **With Context** that is the best combined model trained with a context coming from the classification path of the labels.

We can first observe an improvement of 6.6% in the POS tagging accuracy when using the context, which stresses the importance of such context. However, this only improves the translation accuracy by 2.62%. The improvement in POS tagging does not translate directly into a translation improvement, due to the

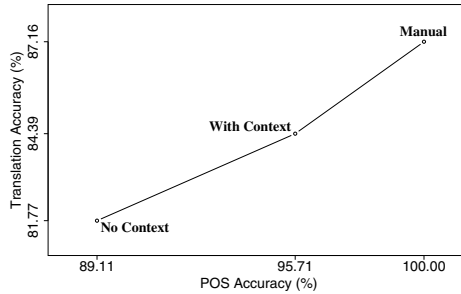


Fig. 3. Contribution of POS accuracy to the translation accuracy

other modules of the pipeline, such as the word sense disambiguation module, whose performance also influences the overall translation accuracy. Indeed, if we evaluate the translation with the manual POS tagging (*Manual* point in Fig. 3), we observe that even with a “perfect” tagging, the translation accuracy does not improve much more. In comparison, a “perfect” tokenization (with a contextless POS tagging), improves the translation accuracy only by 0.02%.

The approach we propose in this paper, with more accurate NLP models and the language structure analysis, achieves an accuracy of 84.39% in this application domain. This is a 17.95% improvement over the state of the art translation approach from [24] that reaches a 66.44% precision.

Analysing the errors, we observe that incorrect recognition of atomic concepts accounts for 22.94% of wrongly translated labels. In the remaining 77.06% of wrongly translated labels the errors are split into two groups: 79.29% due to incorrectly disambiguated senses and 20.71% due to incorrectly recognized formula structure. This suggests directions for further improvements of the approach.

7 Conclusions

We have explored and analysed the natural language metadata represented in several large classifications. Our analysis shows that the natural language used in classifications is different from the one used in normal text and that language processing tools need an adaptation to perform well. We have shown that a standard part-of-speech (POS) tagger could be accurately trained on the specific language of the metadata and that we improve greatly its accuracy compared to the standard long texts models for tagging.

A large scale analysis of the use of POS tags showed that the metadata language is structured in a limited set of patterns that can be used to develop accurate (up to 99.81%) lightweight Backus-Naur form grammars. We can then use parsers based on these grammars to allow a deeper understanding of the metadata semantics. We also show that, for such tasks as translating classifications into lightweight ontologies for use in semantic matching it improves the accuracy of the translation by almost 18%.

References

1. Euzenat, J., Shvaiko, P.: *Ontology Matching*. Springer, Heidelberg (2007)
2. Doan, A., Halevy, A.Y.: Semantic integration research in the database community: A brief survey. *AI Magazine* 26, 83–94 (2005)
3. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Semantic schema matching. In: *Proceedings of CoopIS*, pp. 347–365 (2005)
4. Giunchiglia, F., Yatskevich, M., Avesani, P., Shvaiko, P.: A large dataset for the evaluation of ontology matching systems. *KERJ* 24, 137–157 (2008)
5. Giunchiglia, F., Shvaiko, P., Yatskevich, M.: Discovering missing background knowledge in ontology matching. In: *ECAI*, pp. 382–386. IOS Press, Amsterdam (2006)
6. Zaihrayeu, I., Sun, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X.: From web directories to ontologies: Natural language processing challenges. In: *ISWC/ASWC*, pp. 623–636 (2007)
7. Giunchiglia, F., Soergel, D., Maltese, V., Bertacco, A.: Mapping large-scale knowledge organization systems. In: *ICSD* (2009)
8. Giunchiglia, F., Zaihrayeu, I.: Lightweight ontologies. In: *EoDS*, pp. 1613–1619 (2009)
9. Giunchiglia, F., Zaihrayeu, I., Kharkevich, U.: Formalizing the get-specific document classification algorithm. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) *ECDL 2007*. LNCS, vol. 4675, pp. 26–37. Springer, Heidelberg (2007)
10. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) *ESWC 2009*. LNCS, vol. 5554, pp. 429–444. Springer, Heidelberg (2009)
11. Fuchs, N.E., Kaljurand, K., Schneider, G.: Attempto controlled english meets the challenges of knowledge representation, reasoning, interoperability and user interfaces. In: *FLAIRS Conference*, pp. 664–669 (2006)
12. Schwitter, R., Tilbrook, M.: Lets talk in description logic via controlled natural language. In: *LENLS* (2006)
13. Denaux, R., Dimitrova, V., Cohn, A.G., Dolbear, C., Hart, G.: Rabbit to OWL: Ontology authoring with a CNL-based tool. In: *CNL* (2009)
14. Schwitter, R., Ljungberg, A., Hood, D.: ECOLE — a look-ahead editor for a controlled language. In: *EAMT-CLAW*, pp. 141–150 (2003)
15. Bernstein, A., Kaufmann, E.: GINO — a guided input natural language ontology editor. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) *ISWC 2006*. LNCS, vol. 4273, pp. 144–157. Springer, Heidelberg (2006)
16. Cregan, A., Schwitter, R., Meyer, T.: Sydney OWL syntax — towards a controlled natural language syntax for OWL 1.1. In: *OWLED* (2007)
17. Pool, J.: Can controlled languages scale to the web? In: *CLAW at AMTA* (2006)
18. Wang, C., Xiong, M., Zhou, Q., Yu, Y.: PANTO: A portable natural language interface to ontologies. In: Franconi, E., Kifer, M., May, W. (eds.) *ESWC 2007*. LNCS, vol. 4519, pp. 473–487. Springer, Heidelberg (2007)
19. Fuchs, N.E., Schwitter, R.: Web-annotations for humans and machines. In: Franconi, E., Kifer, M., May, W. (eds.) *ESWC 2007*. LNCS, vol. 4519, pp. 458–472. Springer, Heidelberg (2007)

20. Hepp, M., de Bruijn, J.: GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 129–144. Springer, Heidelberg (2007)
21. Santorini, B.: Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, University of Pennsylvania (1990) (3rd revision, 2nd printing)
22. Morton, T.: Using Semantic Relations to Improve Information Retrieval. PhD thesis, University of Pennsylvania (2005)
23. Kucera, H., Francis, W.N., Carroll, J.B.: Computational Analysis of Present Day American English. Brown University Press (1967)
24. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: algorithms and implementation. In: JoDS, IX (2007)
25. Collins, M.: Head-driven statistical models for natural language parsing. *Computational Linguistics* 29(4), 589–637 (2003)

Measuring Effectiveness of Geographic IR Systems in Digital Libraries Evaluation Framework and Case Study

Damien Palacio¹, Guillaume Cabanac²,
Christian Sallaberry¹, and Gilles Hubert²

¹ Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex

² Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9

Abstract. Common search engines process users' queries (i.e., information needs) by retrieving documents from pre-built term-based indexes. For digital libraries, such approaches are limited regarding particular contexts, such as specialized collections (e.g., cultural heritage collections) or specific retrieval criteria (e.g., multidimensional criteria). In this paper, we consider Information Retrieval systems exploiting geographic dimensions: spatial, temporal, and topical dimensions. Our contribution is twofold as we propose a Geographic Information Retrieval system evaluation framework and test the following hypothesis: combining spatial and temporal dimensions along with the topical dimension improves the effectiveness of Information Retrieval systems.

1 Introduction

Printed literature digitization is currently making significant progress. While some projects only aim to create digital counterparts of physical documents, domain-specific efforts often have more ambitious goals. For example, textual documents are annotated and indexed according to domain-specific models for improving users' experience with document contents [1]. Indeed, Cultural Heritage Libraries generate much digitizing initiatives. The promotion of such collections is then supported by Library Management Systems generally involving full-text Information Retrieval (IR) engines.

In this context, the PIV [2] project, for 'Virtual Itineraries in the Pyrenees Mountains' [2], aims to manage digitized a collection of documents published in the 19th century about the French Pyrenees Mountains. This collection is mainly comprised of newspapers, novels, and travelogues. Local governments foster initiatives aiming at larger dissemination supported by the Web and dedicated IR services. In the meantime, the ratio of geographic queries submitted to usual search engines varies between 12.7% and 18.6% regarding Excite [3], AOL [4],

¹ PIV project is funded by the Pau City Council and the MIDR multimedia library.

and Yahoo [5]. Although query-based common search engines are deemed to deliver accurate results, Kanhabua et al. [6] reported poor precision when it comes to answering geographic queries. As a result, users spend much time skimming the retrieved documents, looking for those that satisfy their information needs. In such a digital library context, it happens that term matching has limitations [7]. For example, the query ‘during the 1810’s’ submitted to a common search engine retrieves only documents containing ‘1810,’ without retrieving ‘1811,’ ‘1812,’ and so on. Similarly, the query ‘Paris’ retrieves documents containing ‘Paris,’ but not ‘Eiffel Tower,’ ‘Louvre Museum,’ and so on. One way of improving systems accuracy is to include the geographic dimension into the retrieval process. We consider the usual acceptance that Geographic Information gathers three dimensions, namely spatial, temporal and, topical [8]. A typical illustration of this is: ‘Fortified towns in South London suburbs during the 13th century.’

Following up recent work on Digital Libraries [7], the main goal of PIV project is to help users finding accurate information inside books. We intend to overcome usual IR Systems (IRSs) limitations regarding geographic information management. Thus, we designed three process chains for indexing spatial [2], temporal [9] and topical [10] information. These allow the retrieval of document units along with associated relevance scores. This work involves various domains, such as Natural Language Processing (NLP), Geographic Information Systems (GISs), Information Retrieval (IR), and Geographic Information Retrieval (GIR). More specifically, this paper tackles both design and experiment of an evaluation framework dedicated to GIR systems. This framework *i*) addresses the evaluation of spatial, temporal and topical IRSs, as well as the evaluation of any combination of these; *ii*) designs a test collection suited to GIR context, which may lead us to implement an original GeocLEF [11]-like track.

The paper is organized as follows. Existing evaluation frameworks dedicated to GIR systems are presented in Sect. 2. In Sect. 3, we propose an evaluation framework addressing GIR systems. Section 4 develops a case study: it presents and evaluates PIV core components—spatial, temporal, and topical IRSs—and their combination. Our hypothesis is: combining these three dimensions is more effective than any of the underlying IRSs. This is tested through an experiment complying with the proposed evaluation framework. In Sect. 5, we review the literature related to GIR systems; these may benefit from our framework. Finally, Sect. 6 concludes the paper and outlines research directions.

2 Suitability of Existing Evaluation Frameworks for GIR

IR has a long tradition of experimentation, especially regarding TREC [12] program dedicated to topical IR evaluation. Moreover, TEMPEVAL [13] evaluation framework is concerned with the temporal dimension. Building on both of these initiatives, Bucher et al. [14] proposed to evaluate two dimensions at the same time: spatial and topical dimensions. This proposal was realized in GeocLEF [11] task of CLEF program [15]. It notably allowed Perea-Ortega et al. [16] to evaluate the effectiveness of classical topical IRSs, such as Terrier [17].

To the best of our knowledge, GIR contributions (reviewed in Sect. 5) were mostly evaluated according to efficiency (e.g, index size and retrieval performance in time). However, it may be worth complementing such quantitative figures with effectiveness evaluation. Moreover, no work to date considered evaluating the three dimensions altogether. Consequently, it is not possible to compare search engines handling these features yet. That is the reason why we propose in the next section an evaluation framework dedicated to GIR.

3 Design of a Framework for GIR Evaluation

The proposed evaluation framework builds on existing state-of-the-art methodologies (especially related to TREC and GeocLEF), and integrates the lacking specificities regarding geographic information. Section 3.1 details the design of a test collection covering the three geographic dimensions; then Sect. 3.2 reports the analysis of PIV GIR system, enabling us to assess its effectiveness.

3.1 Test Collection Supporting GIR Evaluation

In the IR literature, especially at TREC [18], a test collection is comprised of the three following components:

1. A set of n *topics* representing users' information needs. Each topic is at least provided with a title (a keyword-based query), a description (usually a sentence in ordinary language), and a narrative (a detailed explanation of expected information as well as criteria for judging a document as relevant or non-relevant). While a minimum of 25 topics are required for conducting sound statistical analyses [18], note that 50 topics is standard at TREC.
2. The *corpus* comprising numerous documents, some of which are relevant for the proposed topics. A regular TREC corpus for classical ad hoc task represents from 800,000 to 1 million documents [18].
3. The *qrels* (i.e., query relevance judgments) associating each topic with the documents that an individual would expect to retrieve, i.e., a set of relevant documents. Since the corpus is too huge to be extensively considered looking for relevant documents, IR evaluation frameworks rely on the 'pooling' technique, especially at TREC. For each topic t , a document pool is created from the top 100 documents retrieved by the participants' IRs, duplicates being removed. It is hypothesized that resorting to multiple and diverse IRs leads to finding most of the relevant documents belonging to the corpus. Finally, an assessor skims through each document for evaluating whether it matches the information need corresponding to topic t or not: the document is then qualified as *relevant* or *non-relevant*.

Such test collections were operated for several evaluation frameworks, especially at TREC and GeocLEF. Notice that they do not cover all the three dimensions of geographic information. This motivates our work, as we propose to customize the design of test collections in order to enable GIR evaluation by providing:

1. *Topics* covering part or the totality of the three dimensions. For instance, a topic may be titled ‘Potato Famine in Southern Eire after mid-19th century’ and its narrative may be ‘Relevant documents mention scarcity of food and its consequences in Southern Ireland after 1849.’
2. A *corpus* covering the three dimensions: documents present not only the usual topical items but also additional spatial and temporal items.
3. *Qrels* associated with each dimension, resulting from the judgment of relevance between documents and the three dimensions (topical, spatial, and temporal). The co-occurrence of these three dimensions in a given document is not enough for deducing its relevance with respect to the query. Let us consider a document citing ‘Dublin City’ as the protagonist’s place of birth. Although spatially relevant, it does not match the query ‘Pubs in Dublin.’ This subtlety requires the assessment of the global match between the query and the document. Not to overwhelm assessors, we opted for a per dimension binary judgment: a document is either relevant or non-relevant to the considered query and dimension. This rationale is akin to Bucher and colleagues’ conclusions about gradual judgments for each dimension, which were judged as ‘unnecessary cumbersome’ [14]. Finally, considering the three per-dimension binary judgments, as well as the aforementioned global binary judgment, we compute the document relevance value $v \in \{0, 1, 2, 3, 4\}$. This both represents the number of satisfied dimensions, and global relevance. No assumption was made regarding the relative importance of dimensions; they were equitably considered.
4. *Geographic resources* georeferencing spatial entities that occur in the corpus.

The experimental protocol detailed in the next section intends to measure the effectiveness of GIRs. They are evaluated on the basis of the *runs* they provided, i.e., the retrieved document list per topic.

3.2 Protocol for GIRs Comparative Evaluation

The task under evaluation is called ‘ad hoc’ at TREC: an IRS addresses a query by providing a document list ranked by decreasing score. Indeed, the evaluation framework allows effectiveness evaluation for the following IRSs: monodimensional (topical T_o , spatial S , and temporal T_e), bidimensional (T_o+S , T_o+T_e , and $S+T_e$ allowing the measurement of effectiveness improvement according to each missing dimension), and GIRs combining the three dimensions (T_o+S+T_e).

For a given topic, each IRS provides a list of pairs (d, s) representing the score s of each retrieved document d . Usually, effectiveness of an IRS is evaluated with respect to *Average Precision* (*AP*) measure for each topic, and *Mean Average Precision* (*MAP*) overall. These require binary qrels [19, ch. 8]. In the protocol that we propose, however, qrels are gradual in order to represent the three dimensions of geographic information. These two measures are not suitable indeed. We thus used *Normalized Discounted Cumulative Gain* (*NDCG*) [20] relying on gradual relevance judgments; it was notably used at TREC-9 for the

Web task [18]. *NDCG* implements two principles. On the one hand, highly relevant documents ($v \rightarrow 4$ in our case) are more valuable than marginally ($v \rightarrow 1$) relevant documents. On the other hand, a document is all the less valuable that it is ranked low in the result list, because it is rather unlikely that the user reaches this document. Following the example of TREC, we propose two granularity levels for evaluating an IRS: *i*) topic level is represented by *NDCG* while *ii*) the overall level is computed by *MANDCG*: the mean average of the n *NDCG* values, giving the overall effectiveness of an IRS.

For the overall level, the observed differences $\langle m_i^1 - m_j^1, \dots, m_i^n - m_j^n \rangle$ between two systems for the n topics are reported in per cent (of increase or decrease). We denote m_s^t as the value of measure m achieved by system s for topic t . Statistical test significance of the paired observed differences are also reported with respect to significance p -values resulting from Student's paired two-tailed t -test. Although this test theoretically requires a normal data distribution, Hull [21] states that it is robust to violations of this condition. In practical terms, when $p < \alpha$ where $\alpha = 0.05$ the difference between the tested samples is statistically significant; the smaller p -value, the more significant the difference [21].

4 Case Study: PIV GIR System Evaluation

We experiment PIV prototype with our evaluation framework in order to validate the aforementioned hypothesis: combining the three geographic dimensions improves retrieval effectiveness. Therefore, Sect. 4.1 describes this prototype and Sect. 4.2 reports results of its evaluation.

4.1 PIV GIR System

Indexing: Spatial, Temporal and Topical Process Chains. As proposed by Clough et al. [22], we process each of the three geographic dimensions independently. This can be achieved by building several indexes, one per dimension, as advised in [23]. In this way, one can restrain the search on one criterion and easily manage the indexes (e.g., allowing document addition to the corpus). So, our approach processes indexes independently and combines them later on for supporting multidimensional IR. It contributes to GIR field as defined by Jones and Purves [24], as well as GIR in Digital Libraries as defined by Larson [25].

PIV implements three indexing process chains dedicated to textual document processing. Each process chain builds an index. Spatial and temporal process chains are supported by dedicated NLP services. They provide spatial (SF) and temporal/calendar (CF) feature extraction from textual documents and their interpretation: 'River Thames' is annotated as an absolute SF whereas 'North of the River Thames' is annotated as a relative SF—spatial orientation relation [2]. In the same way, 'Spring 1840' is annotated as an absolute CF whereas 'around Spring 1840' is annotated as a relative CF—temporal adjacency relation [9]. So, spatial and temporal indexes result from various stages. The first stage consists in a syntactico-semantic processing sequence [2]: it addresses SF and CF extraction.

This stage is supported by the **LinguaStream** platform [26]. It is mainly comprised of lexical analysis, morpho-syntactic analysis, and syntactico-semantic analysis relying on DCG (Definite Clause Grammar) rules intending to associate one type and one semantics to any extracted SF and CF. The second stage aims at SF and CF interpretation. This uses the symbolic representation of any SF and CF, and operates specific algorithms to calculate approximated numeric representations: spatial geometries are computed for SF [2] and time intervals are computed for CF [9]. Finally, the third stage standardizes resulting indexes. It consists in a spatial, temporal, and topical tiling process that computes the frequency of these spatial, temporal, and topical tiles within texts and weights them [27]. The resulting indexes allow the implementation of state-of-the-art models for relevance score computation based on such spatial, temporal or topical tiles given their occurrence frequencies within documents. This is discussed in the next section, which details such original strategies.

Retrieval: Combination of Result Lists. These indexing strategies may be associated with discrete or continuous scores depending on the overlap ratio between the SF and the tile (the CF and the tile). This enables us to weigh a tile accordingly. We conducted experiments involving IR weighting schemes (TF, TF-IDF, OkapiBM25) along with discrete and continuous frequency computations. TF formula associated with continuous frequency gave the best results in our context [27]. That is the reason why we evaluate combinations of PIV's spatial and temporal retrieval results hereafter. However, as PIV topical process chain is not fully automated yet, we restrain PIV's topical component to the state-of-the-art Terrier full-text IRS.

Each monodimensional IRS is independent: it builds and queries its own index. PIV is supported by these three different monodimensional (source) IRS. Their results are combined in order to constitute a single result list l . Now, Fox and Shaw [28] introduced the CombMNZ combination operator in the IR field; the resulting combined list l gathers together distinct documents retrieved by the source IRSs. Therefore, the similarity s of a document d in l is computed by adding the similarities of d extracted from source IRSs. This sum is balanced by the number of source IRSs that retrieved d . As a result, for any query q , the higher d is ranked in the result lists of the source IRSs (high similarity s between d and q), the more relevant in l it is (i.e., ranked in the top of l). CombMNZ may be compared to a burden of proof, gathering pieces of evidence: documents retrieved by several source IRSs are so many clues enforcing their presumption of relevance. We validated this principle in a quite different context involving combination of the topical and the semantic dimensions [29].

In addition, Lee [30] compared CombMNZ with other operators on TREC test collections, and demonstrated its effectiveness. That is the reason why we experimented with this operator for combining monodimensional IRS results. As every similarity value s computed by source IRSs may belong to a distinct numeric domain, we normalized them within $[0, 1]$ according to: $\text{normalized_similarity} = \frac{\text{unnormalized_similarity} - \text{minimum_similarity}}{\text{maximum_similarity} - \text{minimum_similarity}}$ [30]. So, for query $q = 8$, Fig. 1(a-c) illustrates results retrieved by three IRSs: each result list is comprised of (d_i, s)

pairs where d_i is a document and s is the computed similarity between q and d_i . Combination of these IRSs results is detailed in Fig. 1(d). It shows CombMNZ similarity values and the corresponding computation details. These similarity values are based on the normalized values of IRS sources, cf. Fig. 1(a-c).

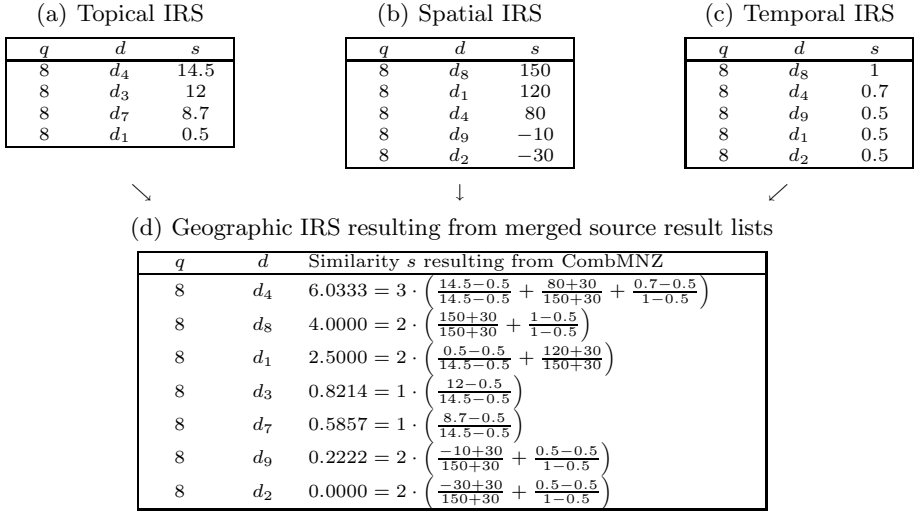


Fig. 1. Illustration of result lists combination with CombMNZ [28]

The example in Fig. 1 illustrates how the score of a d_i takes into account two factors. The more often a source IRS retrieves document d_i , the higher its score s is. Additionally, the higher an IRS ranks d_i in the associated result list, the higher its score s is. In particular, document d_4 illustrates this principle.

Notice that the focus of this paper is on GIRS evaluation. As a result, and due to space limitation, we do not explain PIV components in greater detail. Nevertheless, the interested reader may refer to [29,27] in this respect.

4.2 Evaluation Framework Use for Evaluating the PIV System

We applied the evaluation framework presented in Sect. 3 for evaluating the PIV system. We consider in this section the constituted test collection, the comparative analyses that we carried out, and their limitations.

Design of the MIDR_2010 Test Collection. The MIDR_2010 test collection is comprised of the four components identified in Sect. 3.1. First, the *corpus* collects 5,645 paragraphs extracted from 11 books published between the 18th and 20th centuries, and belonging to Aquitaine Regional Library. They were scanned and processed with OCR software. A document d , as retrieved by the IRS, is one of these paragraphs; it is considered as the best entry point in its associated book.

Second, 31 *topics* covering the three dimensions of geographic information were constituted. Third, *qrels* were obtained by querying three IRS—a topical IRS based on PL2 IR model (built-in Terrier configuration), a spatial IRS and a temporal IRS—with the ‘title’ part of the topics. Lastly, the *geographic resources* corresponding to the corpus are provided by the French National Geographic Institute (BD NYME ® database). For each topic, the results retrieved by the IRSs were considered for setting up the pool. It was then assessed according to binary judgments for each dimension, and for the global judgment. These four judgments were aggregated in a single gradual value, as presented in Sect. 3.1.

Comparative Analysis of IRS Effectiveness. In Tab. 1, we report the observed comparisons between various IRSs and two baselines identified in [16]: To^+ is a strong baseline corresponding to OkapiBM25 model; To^- is a weaker baseline corresponding to TF-IDF model. In addition, S denotes the spatial IRS, and Te denotes the temporal IRS. The reported results show effectiveness of search engines according to the 31 tested topics. Overall, the two baselines showed similar effectiveness. Contrary to the observations presented in [16], TF-IDF achieves better performance than OkapiBM25 in our experiment. This difference may be due to the fact that the MIDR_2010 test collection is comprised of document paragraphs similar in length, as opposed to plain documents with variable lengths, for which OkapiBM25 is known to achieve best performance.

Table 1. IRS effectiveness w.r.t. topical baselines. The * symbol ([†] symbol) denotes a significant difference compared with baseline To^- (baseline To^+).

Combination of N IRSs	Monodimensional IRS				$MANDCG$	Gain (%)	
	To^-	To^+	S	Te		To^-	To^+
1	✓				0.4726	0.0	0.1
		✓			0.4721	-0.1	0.0
			✓		0.4574	-3.2	-3.1
				✓	0.4836	2.3	2.4
2	✓	✓			0.4722	-0.1	0.0
	✓		✓		0.6162* [†]	30.4	30.5
	✓			✓	0.7017* [†]	48.5	48.6
		✓	✓		0.6165* [†]	30.4	30.6
			✓	✓	0.7017* [†]	48.5	48.6
				✓	0.6993* [†]	48.0	48.1
3	✓	✓	✓		0.6104* [†]	29.2	29.3
	✓	✓		✓	0.6842* [†]	44.8	44.9
	✓		✓	✓	0.7852* [†]	66.1	66.3
		✓	✓	✓	0.7859 * [†]	66.3	66.5
4	✓	✓	✓	✓	0.7578* [†]	60.3	60.5

Regarding monodimensional IRSs they all achieve similar performance; best effectiveness (0.4836) is reached by the temporal IRS. In addition, combining at least two heterogeneous dimensions yields better performance. Notice that the associated improvement is statistically significant regarding the two baselines,

$\text{Te}+\text{To}^+$ and $\text{Te}+\text{To}^-$ being the most effective combinations (0.7017). However, the combination $\text{Te}+\text{S}$ is similar in effectiveness (0.6993). An explanation for this may involve absolute spatial entities (e.g., ‘Paris’), which are easily retrieved by a topical IRS (exact match). However, only a spatial IRS can properly process more complex queries involving relative spatial entities (e.g., ‘Eastern Paris’).

Combining the three dimensions (0.7859) yields better results (+12.0%) than the best bidimensional combination (0.7017), which is statistically significant ($p = 0.000$). Adding to these dimensions a topical IRS (To^+ or To^-) does not result in more improvement (0.7578). The resulting topical reinforcement may lessen the complementary information contributed by the two other dimensions. In conclusion, combining the three dimensions provides the best performance (0.7859). The 66.3% improvement regarding To^- validates the hypothesis formulated in this paper: the combination of the three geographic information dimensions yields better performance than considering only the topical dimension.

Limitations of the Current Evaluation. The experiment reported in this paper presents at least two limitations. On the one hand, comprising 5,645 paragraphs for a 3.7 Mb total size, the MIDR_2010 test collection is very limited in size compared with TREC collections. On the other hand, the experiment was completed with 31 topics; this represents six more topics than the minimum number of topics to use for conducting proper statistical analysis. We keep on judging documents manually in order to provide more topics.

Despite these limitations, the evaluation framework proposed in this paper is appropriate for experimenting with the various proposals found in the GIR domain. The next section briefly introduces a representative sample of this work.

5 GIR Related Work

Work related to GIR includes the following prominent five projects. GIPSY [31], for ‘Georeferenced Information Processing System,’ proposes a method for indexing textual documents; it is based on the aggregation of the footprints corresponding to spatial entities. This aggregation is used to find the most representative geographic areas in order to index a document. GeoSem [26], for ‘Geographic Semantic,’ is dedicated to document (texts, maps, charts) geographic information semantics processing. SPIRIT [32], for ‘Spatially-Aware Information Retrieval on the Internet,’ aims to find Web pages that refer to places or geographic areas specified in a query. STEWARD [33], for ‘Spatio-Textual Extraction on the Web Aiding Retrieval of Documents’, performs extraction, retrieval, and geographic area visualization for unstructured texts. CITER [34] for ‘Creation of a European History Textbook Repository’ offers history textbook retrieval according to several dimensions. Finally, DIGMAP [35], for ‘Discovering our Past World with Digitised Maps,’ is dedicated to cultural and scientific heritage promotion. These geographic information indexing and retrieval systems handle in priority the spatial criteria (SF, cf. Sect. 4.1). They all use pre-built monodimensional indexes and propose similar approaches to merge result lists. They apply

filtering-like approaches: for instance, STEWARD retrieves topical relevant document units first, and goes on filtering out these results according to the spatial dimension. This is quite different from the CombMNZ [28] combination-based approach that we introduced in this paper.

6 Conclusion and Future Work

We considered geographic IRSs handling spatial, temporal, and topical dimensions. However, common search engines show limitations in such contexts. Consequently, our contribution is twofold: we propose an evaluation framework, and use it for validating our hypothesis: combining the three dimensions improves the accuracy of retrieval results. Applying this framework on an appropriate test collection showed an improvement of 66.3% over a topical baseline. Moreover, this performance gain is statistically significant. These good results give an empirical validation of our proposals experimented with the PIV GIR system [2]. In addition, this evaluation framework is not restricted to the three mentioned dimensions: it can also integrate other dimensions, such as confidence in the information, and its freshness [36].

In addition to experimenting with a larger test collection, we are now intending to propose and experiment alternate IRSs combination approaches. We are especially interested in constraint-based combination methods—involving concepts of requirement and preference—based on a linear approach [37], or related to fuzzy OWA [38]-based approaches. Having demonstrated the feasibility of GIR evaluation in this paper, we plan to propose this framework in a GeocLEF-like track. Its main originality is to provide documents, and allow the evaluation of IRSs according to the three dimensions of geographic information (i.e., topical, spatial and temporal). For this purpose, the constituted MIDR_2010 test collection is available on PIV project website².

References

1. Sautter, G., Böhm, K., Padberg, F., Tichy, W.F.: Empirical Evaluation of Semi-automated XML Annotation of Text Documents with the GoldenGATE Editor. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 357–367. Springer, Heidelberg (2007)
2. Gaio, M., Sallaberry, C., Etcheverry, P., Marquesuzaa, C., Lesbegueries, J.: A global process to access documents' contents from a geographical point of view. *J. Vis. Lang. Comput.* 19(1), 3–23 (2008)
3. Sanderson, M., Kohler, J.: Analyzing Geographic Queries. In: SIGIR-GIR 2004: Workshop on Geographic Information Retrieval at SIGIR (2004)
4. Gan, Q., Attenberg, J., Markowetz, A., Suel, T.: Analysis of geographic queries in a search engine log. In: *LocWeb 2008: 1st Int. Workshop on Location and the Web*, pp. 49–56. ACM, New York (2008)
5. Jones, R., Zhang, W.V., Rey, B., Jhala, P., Stipp, E.: Geographic intention and modification in web search. *Int. J. Geogr. Inf. Sci.* 22(3), 229–246 (2008)

² <http://t2i.univ-pau.fr/MIDR/>

6. Kanhabua, N., Nørvåg, K.: Temporal Language Models for Determining Time of Non-timestamped Documents. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 358–370. Springer, Heidelberg (2008)
7. Liesaputra, V., Witten, I.H., Bainbridge, D.: Searching in a Book. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 442–446. Springer, Heidelberg (2009)
8. Usery, E.L.: A feature-based geographic information system model. *Photogramm. Eng. Rem. Sens.* 62(7), 833–838 (1996)
9. Le Parc-Lacayrelle, A., Gaio, M., Sallaberry, C.: La composante temps dans l'information géographique textuelle. *Document Numérique* 10(2), 129–148 (2007)
10. Sallaberry, C., Baziz, M., Lesbegueries, J., Gaio, M.: Towards an IE and IR System Dealing with Spatial Information in Digital Libraries – Evaluation Case Study. In: ICEIS 2007: 9th Int. Conference on Enterprise Information Systems, pp. 190–197 (2007)
11. Gey, F.C., Larson, R.R., Sanderson, M., Joho, H., Clough, P., Petras, V.: GeoCLEF 2005: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 908–919. Springer, Heidelberg (2006)
12. Voorhees, E.M., Harman, D.K.: *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (2005)
13. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., Pustejovsky, J.: The TempEval challenge: identifying temporal relations in text. *Lang. Resour. Eval.* 43(2), 161–179 (2009)
14. Bucher, B., Clough, P., Joho, H., Purves, R., Syed, A.K.: Geographic IR Systems: Requirements and Evaluation. In: ICC 2005: 22nd Int. Cartographic Conference (2005) (CDROM)
15. Peters, C.: Introduction. In: Peters, C. (ed.) CLEF 2000. LNCS, vol. 2069, pp. 1–6. Springer, Heidelberg (2001)
16. Perea-Ortega, J.M., García-Cumbreras, M.A., García-Vega, M., Ureña-López, L.A.: Comparing Several Textual Information Retrieval Systems for the Geographical Information Retrieval Task. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (eds.) NLDB 2008. LNCS, vol. 5039, pp. 142–147. Springer, Heidelberg (2008)
17. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier Information Retrieval Platform. In: Losada, D.E., Fernández-Luna, J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 517–519. Springer, Heidelberg (2005)
18. Harman, D.K.: The TREC Test Collections. In: [12], ch. 2, pp. 21–53
19. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (July 2008)
20. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.* 20(4), 422–446 (2002)
21. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In: SIGIR 1993: 16th Annual Int. SIGIR Conference, pp. 329–338. ACM Press, New York (1993)
22. Clough, P., Joho, H., Purves, R.: Judging the Spatial Relevance of Documents for GIR. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsirikika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 548–552. Springer, Heidelberg (2006)

23. Martins, B., Silva, M.J., Andrade, L.: Indexing and ranking in Geo-IR systems. In: GIR 2005: Workshop on Geographic Information Retrieval, pp. 31–34. ACM, New York (2005)
24. Jones, C.B., Purves, R.: GIR 2005 ACM Workshop on Geographical Information Retrieval. SIGIR Forum 40(1), 34–37 (2006)
25. Larson, R.R.: Geographic Information Retrieval and Digital Libraries. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 461–464. Springer, Heidelberg (2009)
26. Bilhaut, F., Charnois, T., Enjalbert, P., Mathet, Y.: Geographic reference analysis for geographic document querying. In: HLT-NAACL 2003: Workshop on Analysis of Geographic References, pp. 55–62. ACL, Morristown (2003)
27. Palacio, D., Sallaberry, C., Gaió, M.: Normalizing Spatial Information to Improve Geographical Information Indexing and Retrieval in Digital Libraries. In: ISGIS 2010: Joint Int. Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science Proceedings (to appear, 2010)
28. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: Harman, D.K. (ed.) TREC-1: 1st Text REtrieval Conference, Gaithersburg, MD, USA, pp. 243–252. NIST (February 1993)
29. Hubert, G., Mothe, J.: An adaptable search engine for multimodal information retrieval. J. Am. Soc. Inf. Sci. Technol. 60(8), 1625–1634 (2009)
30. Lee, J.H.: Analyses of Multiple Evidence Combination. In: SIGIR 1997: 20th Annual Int. SIGIR Conference, pp. 267–276. ACM Press, New York (1997)
31. Woodruff, A.G., Plaunt, C.: Gypsy: automated geographic indexing of text documents. J. Am. Soc. Inf. Sci. 45(9), 645–655 (1994)
32. Vaid, S., Jones, C.B., Joho, H., Sanderson, M.: Spatio-textual Indexing for Geographical Search on the Web. In: Bauzer Medeiros, C., Egenhofer, M.J., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 218–235. Springer, Heidelberg (2005)
33. Lieberman, M.D., Samet, H., Sankaranarayanan, J., Sperling, J.: STEWARD: Architecture of a Spatio-Textual Search Engine. In: GIS 2007: 15th Annual ACM Int. Symposium on Advances in Geographic Information Systems, pp. 1–8. ACM, New York (2007)
34. Pfoser, D., Efentakis, A., Hadzilacos, T., Karagiorgou, S., Vasiliou, G.: Providing Universal Access to History Textbooks: A Modified GIS Case. In: Carswell, J.D., Fotheringham, A.S., McArdle, G. (eds.) W2GIS 2009. LNCS, vol. 5886, pp. 87–102. Springer, Heidelberg (2009)
35. Manguinhas, H., Martins, B., Borbinha, J., Siabato, W.: The DIGMAP Geo-Temporal Web Gazetteer Gervice. e-Perimtron: Int. Web J. Sci. Technol. Affined Hist. Cartogr. Maps 4(1), 9–24 (2009)
36. Costa Pereira, C., Dragoni, M., Pasi, G.: Multidimensional relevance: A new aggregation criterion. In: ECIR 2009: 31th European Conference on IR Research on Advances in Information Retrieval, pp. 264–275. Springer, Heidelberg (2009)
37. Farah, M., Vanderpooten, D.: An outranking approach for information retrieval. Inf. Retr. 11(4), 315–334 (2008)
38. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans. Syst. Man Cybern. 18(1), 183–190 (1988)

A Visual Digital Library Approach for Time-Oriented Scientific Primary Data

Jürgen Bernard¹, Jan Brase², Dieter Fellner^{1,3},
Oliver Koepler², Jörn Kohlhammer^{1,3}, Tobias Ruppert³,
Tobias Schreck¹, and Irina Sens²

¹ Technische Universität Darmstadt, Germany

² German National Library of Science and Technology, Hannover, Germany

³ Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany

Abstract. Digital Library support for textual and certain types of non-textual documents has significantly advanced over the last years. While Digital Library support implies many aspects along the whole library workflow model, interactive and visual retrieval allowing effective query formulation and result presentation are important functions. Recently, new kinds of non-textual documents which merit Digital Library support, but yet cannot be accommodated by existing Digital Library technology, have come into focus. Scientific primary data, as produced for example, by scientific experimentation, earth observation, or simulation, is such a data type. We report on a concept and first implementation of Digital Library functionality, supporting visual retrieval and exploration in a specific important class of scientific primary data, namely, time-oriented data. The approach is developed in an interdisciplinary effort by experts from the library, natural sciences, and visual analytics communities. In addition to presenting the concept and discussing relevant challenges, we present results from a first implementation of our approach as applied on a real-world scientific primary data set.

Keywords: Visual Analysis, Visual Search, Content-Based Search, Scientific Primary Data, Visual Cluster Analysis.

1 Introduction

Digital Library systems are indispensable elements of an effective information infrastructure. Modern acquisition, processing, storage, and delivery technologies have improved existing and created totally new ways by which libraries can serve users. For example, Web technologies enable distributed user access; full text processing allows issuing specific, on-target queries and services may be enhanced by recommendation and personalization functionality. While much of this functionality is available in existing Digital Library systems, it is most often restricted to *textual* documents. While text is of high importance, increasingly, *non-textual* document types arise in many application areas and treating these with library services is desirable. This is quite obvious for popular non-textual

document types such as digital image, video, and audio content. In these cases, results from Multimedia Processing and Retrieval apply and can be used to realize content-based search and presentation for such content.

While ubiquitous and relevant, such multimedia document types are not the only, nor the per se most important document types. In recent discussion among research institutions and research funding agencies [12], *scientific primary data* has been identified as a document type worth considering strategically. Consequently, development of infrastructure to support indexing, storage, accessing, delivery, and archival of scientific primary data is identified as a necessity. Let two out of many relevant observations motivate this point. (a) *Re-usage* of scientific data is desirable to increase transparency of research and research results, and to lower the cost by sharing of data; and (b) *archival* of scientific primary data is useful for possible re-examination of that data in the future, when new analysis methods may become available. Consider *climate data* for an example, which is expensive to obtain (involving large scale and distributed observation). In the future, novel climate analysis programs may become available, where historic data can support calculation of more accurate climate models. Library support for such data clearly would benefit science and society.

For illustration purposes we describe a possible application scenario for our system. Here, a natural scientist detects an interesting *curve progression* in her collected measurements. According to her hypothesis, this exemplary time series pattern might indicate a future event that is relevant to her research. To verify the hypothesis that there is a connection between her measurements and the event, she wants to examine similar curve progressions in related data sets. A requirement for this task is a visual overview of the most similar data sets grouped by their similarity to the chosen reference example. Furthermore, measurements in the same category (e.g. global radiation) are a matter of particular interest. This is obtained by offering filtering options that operate on the meta-information appended to the data. Besides defining a search pattern by choosing a curve progression example from the existing data (“query-by-example”), a scientist wants to search for an artificial curve sketched manually (“query-by-sketch”). This can be realized in a visual-interactive graphical interface. Finally, the results of the scientist’s query are displayed in the same time scale to analyze correlations between the detected time series.

Devising and implementing Digital Library support for tasks like the above mentioned is a complex challenge that involves finding solutions on many levels, ranging from acquisition to standardization over to retrieval, delivery, and archiving. In this work, we focus on the specific problem of visual retrieval and exploration in large sets of *time-oriented* scientific primary data, as an important subtype of scientific primary data in general. We present a concept devised as well as early results developed in the course of a joint research project carried out by librarians, computer scientists, and natural scientists. Our approach adapts and combines techniques from time series analysis, multimedia retrieval, and information visualization, and will be prototypically implemented

and evaluated in practice. The results presented are one step towards advanced Digital Library support for this kind of data.

2 Background and Related Work

We review related work in Digital Libraries, scientific primary data initiatives, and retrieval and visualization in time series data.

2.1 Scientific Primary Data in the Digital Library Context

Digital Library systems have evolved over time from purely academic and pioneering works, to standardized and established systems, which are available for practical usage. Popular example systems include Fedora [3], Greenstone [4], and DLib [5]. These systems typically are oriented towards textual documents, considering non-textual documents as uninterpreted digital content for which no native system support is provided. Digital Library systems for non-textual documents which allow content-based search are relatively scarce in practice, owing to the high variability between and within collections of non-textual documents making standardization difficult. Prototypical systems exist for a number of multimedia document types, including music [6] or image and other multimedia documents [7]. These systems offer advanced support for indexing and visual retrieval of certain content. For example, the PROBADO3D system [8] supports searching in architectural model data by means of global shape and room structure, and allows for visual queries specification.

Scientific primary data may also be regarded as a non-textual document type. It often comprises numeric data or georeferenced data on continuous or discrete scales and stem from many different sources including earth observation, experimentation, or simulation. The primary data is usually also associated with textual metadata including data description, author and origin information, and even references to corresponding publications. While the necessity of treating scientific primary data by library services is generally recognized, significant challenges exist to this end including [1] but not limited to (a) persistent storage of massive volumes of data; (b) standardization of data formats and encoding; (c) quality control, peer review, and citability of data sets; and (d) clarification of legal aspects regarding ownership, access, and re-usage.

To date, a number of operational Digital Library systems for scientific primary data already exist. Examples include PsychData [9] (psychological data), PANGAEA [10] (geoscientific and environmental data), or Drayd [11] (generic data underlying natural sciences publications). Several research projects address conceptual challenges and implications in this area. The KoLaWiss initiative [2] identified organizational, technical, economic and data type-oriented challenges for establishing a collaborative scientific primary data infrastructure. Citability and publication of this data has been devised by the project “Publication and Citation of Scientific Primary Data” [12]. Establishing the European infrastructure for biological information is aimed at by the ELIXIR [13] coordination research initiative. Approaches towards service-oriented infrastructure in the Arts and Humanities are considered in the project BAMBOO [14].

2.2 Search and Visualization in Time-Oriented Data

As denoted by the example in the introduction chapter, content-based access to time series data requires the definition of similarity measures, which is important for search and visual clustering purposes. Liao [15] surveys many measures for time series similarity estimation, distinguishing three groups of time series similarity calculation approaches: raw data-based, model-based, and feature-based. Raw data-based (or transformation) approaches directly compare time series raw data, usually by measuring the cost of transforming one series to match another [16]. Model-based approaches work by calculating the degree to which two time series to be compared share the same underlying statistical model. In the feature vector (or descriptor) approach, descriptor metadata is automatically extracted from the time series data. Then, the similarity between two time series is estimated by the distance calculated between their respective descriptors. Consequently, the definition of the descriptor extraction algorithm determines the similarity concept. Examples of time series feature extractors rely e.g., on Fourier analysis [17], or on aggregation or discretization approaches [18]. Descriptor approaches usually are robust, amenable to database indexing, and simple to implement. An important conceptual distinction in time series similarity search is between global and partial search. While in global search whole time series are compared, partial search identifies similar subsequences. Techniques for partial similarity search are typically based on Sliding-Windows approaches, or on segmentation approaches such as top-down or bottom-up analysis.

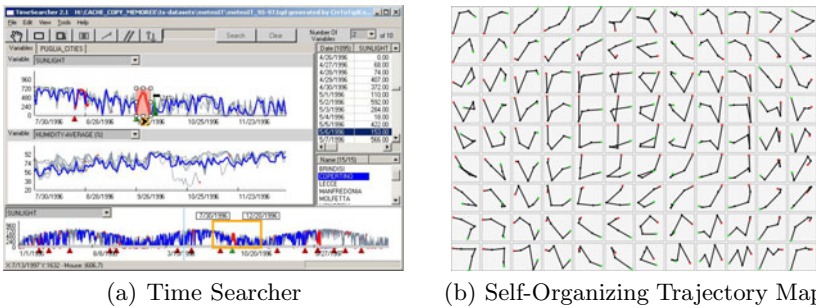


Fig. 1. (a) Time Searcher system [19]. (b) Self-Organizing Map computed for trajectory-oriented data [20].

Searching in time series data can effectively be supported by visual interactive query specification and result visualization. The Time Searcher System [19] (cf. Figure 1(a)) enables interactive query specification via visual filters called Timebox Widgets. These filters define ranges in the time and parameter axis. Similar time series within these ranges are found and highlighted, giving immediate feedback upon query specification. In previous work, we implemented a system for visual exploration of 2D time-dependent scatter data interpreted as trajectory data [20]. Based on a simple geometric descriptor, the system clusters

large sets of trajectory data by means of the Self-Organizing Map algorithm [21]. An early application of SOM to visualization of stock market chart data was explored in [22]. The Self-Organizing Map approach is a popular method for visual cluster analysis due to producing similarity-preserving layouts (cf. Figure II(b) for an illustration). The SOM approach is well-suited to support visual search as a sort of *visual catalog*. Our proposed approach will rely on this algorithm (cf. also Section 4.3).

3 Library-Oriented Treatment of Scientific Primary Data

Recognizing the need for data sharing, several scientific communities have already organized data collection, archiving and access, to serve their community demands. For example, earth and environmental studies data are collected and shared on a worldwide level through the World Data Center System [23]. Data publication is an essential component of every large scientific instrument project (e.g., the CERN Large Hadron Collider). These trends induce development of new library services. DOI-based data set registration and portal-based access are two practical developments in current library support for primary data.

Data set Registration. Data set identification is a key element for citation and long term integration of data sets into text as well as supporting a variety of data management activities. To achieve the rank of a publication, a data publication needs to meet the two main criteria, *persistence* and *quality*. Quality is a rather difficult concept typically addressed by data curators building on domain-dependent guidelines and best practices. Data persistence is a rather technical problem, and addressed by the data hosting infrastructure. Technical infrastructure for data set identification is already practically provided. E.g., the German National Library of Science and Technology (TIB) developed and promotes the use of Digital Object Identifiers (DOI) for data sets. DOI names are already widely used in scientific publishing to cite journal articles. Since 2005, TIB is an official DOI registration agency with a focus on the registration of scientific primary data. In cooperation with several World Data Centers, data collected from various scientific disciplines amounting to over 700,000 data sets have been registered by TIB with DOI names as persistent identifiers.

Portal-Based Access to Remotely Stored Data. Having a DOI-based index of scientific primary data in principle allows the creation of user-friendly portal solutions to browse and access the data, based on textual metadata. An example is the *Get-Info* portal operated by TIB. It bundles access to subject databases, publishing house offerings and library catalogs with integrated full text delivery. The aim is to include all sorts of non-textual information into GetInfo. Primary research data sets are already integrated into GetInfo, and can currently be accessed by metadata queries. The concept presented in this paper is one step toward extending the access to visual and content-based methods for this data sets.

4 Approach and First Results

We describe our concept for visual retrieval in time-dependent scientific primary data and apply it to a concrete scientific data set. The described system forms the baseline for subsequent refinement of search and navigation functionality to be developed in collaboration with scientific users (cf. Section 5).

4.1 Considered Data Set

For initial development we use data from the scientific data information system PANGAEA [10] hosted by the Alfred-Wegener-Institute for Polar and Marine Research, Bremerhaven, Germany and the Center for Marine Environmental Sciences, Bremen, Germany. PANGAEA archives, publishes, and distributes geo-referenced scientific observation data. The data is organized by categories comprising observations on water (e.g., temperature, salinity, oxygen), sediment (e.g., total organic carbon (TOC)), ice (e.g., chemical composition, dust concentration), and atmosphere (e.g., temperature, humidity). Most data sets can be downloaded as text files including the measurement data and accompanying metadata. The latter covers information on citations, originating project name, spatial and temporal conditions, parameter description, etc. The raw data is provided in ASCII table format, containing time stamp and respective measurement data. We are currently considering the raw data, while we will include also metadata (if available) for filtering and query refinement. Our sample data pool consists of over 12,000 data files from the years 1981 to 2009, provided by the project BSRN [24]. The data tables have up to 100 columns corresponding to the number of time series, and a maximum of 50,000 rows, regarding to the number of measured time samples.

The data set is chosen for initial test phases, since it is enriched by a structured meta-information block (see Table 1), which is currently neglected. Future work will address the integration of content-based and meta-information search.

Table 1. Excerpt of meta-information in PANGAEA data files

Meta-Information	Description
Citation	Data set citation (name of author, name of data set, institution, publication year, DOI-Code)
Project	Project name, link to project website
Coverage	Spatial and temporal conditions (time start and end, longitudinal and lateral coordinates, height above sea level)
Event	Description of measurement event (e.g. measurement setup)
Other version	Link to related measurements
Comment	Additional comments
Parameter	Description of parameters, unit, methodology, investigator
Size	Number of rows

4.2 Feature-Based Descriptor Extraction

In our baseline system, we currently support descriptor-based global similarity search in time series, based on the notion of geometric similarity of respective curves. We compute descriptors by application of a work-in-progress modular descriptor calculation pipeline described next (cf. Figure 2).

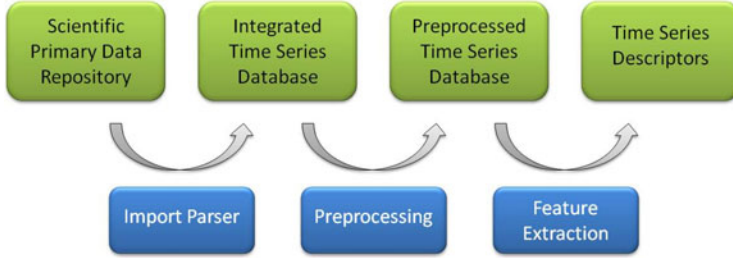


Fig. 2. Feature extraction pipeline

The initial step reads the primary data from provided data files, or from a data repository. Currently we focus on parsing data files from the PANGAEA platform. However, importing time series data from other sources is possible by using dedicated data parsers. After data import, time series preprocessing may be applied as required by the descriptor extraction approach, the application need, and/or condition of the primary data. Several standard normalization techniques including data discretization, transformation, interpolation, and outlier and missing value treatment are implemented and can be applied prior to feature extraction. We currently consider whole time series. However, work is ongoing to implement time series segmentation to support local similarity search as well. After preprocessing, the feature extraction step can basically rely on any appropriate feature proposed so far [15]. Features based on Fourier transform or on discrete approximation have shown to be effective in the literature, and should be supported as baseline similarity functions in our system. For our first experiments, we apply a simple aggregation-based descriptor to reduce each series to a comparable, discretized representation of constant length, which will be used for subsequent clustering and retrieval steps. Considering that the implementation of the descriptor is of utmost importance for the supported similarity concept, the question arises which descriptor and preprocessing options should be chosen for a given search. This is an important research problem relating to the semantic gap, which can be tackled by user evaluation. Our goal is to let the user flexibly select the used descriptors and processing options, finding the best settings for conducting the visual search. Also, techniques based on relevance feedback (RF) are in principle applicable to mediate the semantic gap problem. Addressing interactive and visual descriptor choice is an important aspect of future work in our project.

4.3 A Visual Catalog of Time Series Data for Data Exploration

As our approach suggests an explorative content-based search, we adhere to Shneiderman's *Information Visualization Mantra* [25] ("overview first, zoom and filter, then details on demand"). To create a useful overview for thousands of time series, we propose to offer a "visual catalog" supporting effective data exploration. Two properties of such a catalog we deem useful include (1) reflectance of similarity relations between series data elements for intuitive navigation, and (2) reduction of the data cardinality while identifying the most prominent patterns in the data set. Regarding (1), the patterns should be arranged on the visual display as intuitively as possible. A global ordering of the displayed time series patterns is desirable. The more samples are presented in a sorted way, the better will be the applicator's comprehension. Regarding (2), an appropriate clustering algorithm needs to be applied, which supports (1) and is compatible to the available data descriptors.

After careful consideration, and based on good experience on other data domains, we decided to apply the Self-Organizing Map (SOM) algorithm [21], which addresses the aforementioned requirements. The algorithm is widely used in the clustering domain and has beneficial visualization properties. It is able to reduce a large data set to user-settable number of clusters that are arranged in a low-dimensional grid in an approximately topology-preserving way. For details, we refer to [21]. We apply the SOM approach on a subset of the PANGAEA content, based on our first descriptor implementation. Figure 3 gives an illustration of a SOM map showing a number of clusters of time series patterns from the data set. Applying the example from the introduction, it can be seen in Figure 3 that the natural scientist can obtain an effective overview of the curve shapes of the scientific primary data pool (left image). Furthermore she can pick an example pattern and search the data set for details, which can be displayed on demand (right image).

We consider the SOM approach in combination with an appropriate descriptor as a good candidate for a visual catalog of time series. Based on the overview

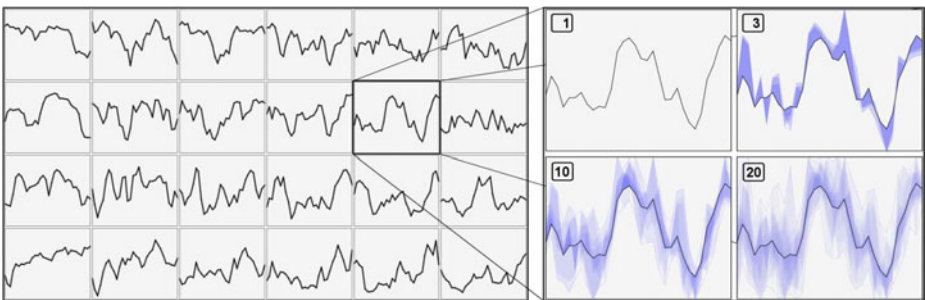


Fig. 3. Left: Visual time series catalog, provided by SOM clustering. Each cell shows one data cluster by a representative time series. Right: A detail view of a selected cluster is shown by an opacity-based overlaying view.

provided by SOM, search interfaces and detail visualization displays can be implemented to support drill-down by the user.

4.4 Visual Query Specification

The visual time series catalog gives an overview of the whole data library. Based on this, content-based user queries may be executed via visual query specification. We initially support two search modalities. A query curve can be specified either by selection from the visual catalog, or by drawing a curve sketch, as described in the use case in Section 4.1. Based on the time series descriptors, distances are calculated, a ranking is obtained, and the results are highlighted by color coding on the catalog itself, or displayed in a separate list view. Figure 4 illustrates.

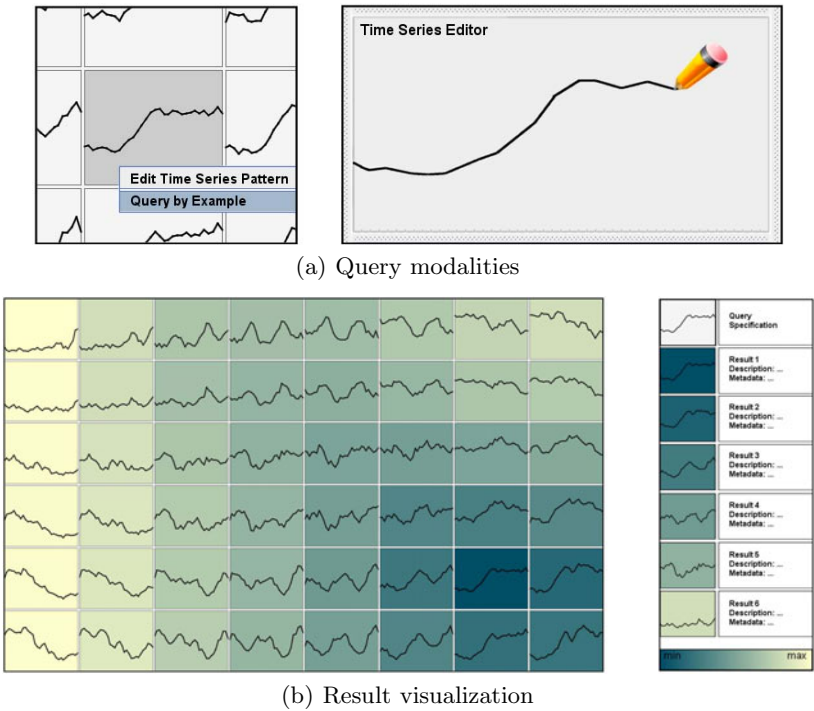


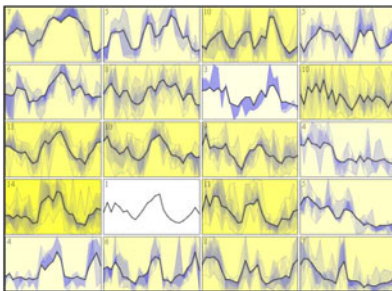
Fig. 4. (a) Query-by-example based on selection and sketching. (b) Result visualization based on the catalog and list.

4.5 Metadata and Export to User Tools

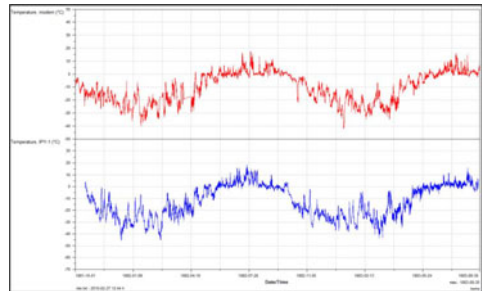
As already indicated, scientific primary data sets are often enriched by meta-information regarding author, originating project, measurement specifics and so on. Of course, such information (if available) must not be neglected in the visual

search. We currently support a light-weight approach to include metadata search. Specifically, uninterpreted full text search in the metadata fields is provided. We point out that metadata integration over heterogeneous data sources is a difficult and expensive process. As we aim to search over heterogeneous data sources, this is a pragmatic approach. In our implementation, a simple text input field enables the user to search in the meta-information of the data sets and filter meaningful time series plots. For example, if the user only wants to consider measurements of a certain researcher she is able to specify her search by typing the researcher's name in a projected meta-information search window. As a result, the data sets authored by the special researcher will be highlighted in the visual catalog (cf. Figure 5(a) for an example).

As our system is intended to support the data-oriented scientific research process, it is important to support domain-dependent tools for export of search results. As a starting point, export of found time series to PanPlot [26], which visualizes small amounts of time series in publication-ready quality, is possible (cf. Figure 5(b) for an example).



(a) Keyword frequency highlighting



(b) Search result export

Fig. 5. (a) Highlighting the frequency of occurring keywords from a metadata search. (b) shows a time series search result exported to a specialized analysis tool (PanPlot).

5 Discussion and Next Steps

Our first step towards visual search in a Digital Library system for time-oriented data is based on the concepts of visual catalog and on content-based queries. Our implemented descriptor supports the similarity notion of global curve shape and is only a starting point. Technically, a wealth of further functionality to explore exists, including design of additional curve shape descriptors, partial similarity, and time- and scale invariant search modalities. We recognize that for the prototype to be successful, it needs to solve real user problems and therefore, further development will take place in close collaboration with scientific users. During an evaluation workshop, we will demonstrate the currently existing prototype to scientists, expecting that relevant use-cases will be defined that can in another iteration be supported in the prototype.

We expect that the most useful search functionalities will not consist of only a single modality (e.g., curve shape), but rather a combination thereof. Additional modalities may involve correlation-based comparison of time series at different scales and possibly applying on a partial level. We further expect that metadata will play an important role, either for filtering of search results or as input to adaptive search algorithms. Conceptually, we are interested in more closely combining browsing and searching. Tight coupling of browsing and searching is expected to yield effective search results. Also, implications regarding scientific data infrastructure are given. For our methods to be broadly applicable, our system needs to interface with many data providers, raising the question of interoperability.

6 Conclusions

We introduced the problem of Visual Digital Library support for scientific primary data. We argued that this data is requiring library support, and that a user-interface based on visual search is desirable. Specifically, content-based visual search should complement purely metadata based search to be effective. A design and development methodology based on visual cataloging and content-based searching in time-oriented data was presented. A first implementation was applied on real data. Options for future work and a user-in-the-loop development model were presented.

Acknowledgments

Rainer Sieger and Hannes Grobe of the Alfred Wegener Institute kindly provided data and initial expert feedback. Tatiana von Landesberger and Sebastian Bremm of Interactive-Graphics Systems Group at TU Darmstadt provided helpful discussion and suggestions. This work was supported by a grant from the Leibniz Association as part of the "Joint Initiative for Research and Innovation" program.

References

1. German Research Foundation (DFG): Report on round table meeting of research data. Whitepaper (January 2008), http://www.dfg.de/download/pdf/foerderung/programme/lis/forschungsprimaerdaten_0108.pdf (in German)
2. Society for Scientific Data Processing Goettingen: Cooperative long-term preservation for research centers. Project Report (April 2009) (in German)
3. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.* 6(2), 124–138 (2006)
4. Witten, I.H., McNab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: *Proceedings of the Fifth ACM International Conference on Digital Libraries* (2000)
5. Castelli, D., Pagano, P.: Opendlib: A digital library service system. In: Agosti, M., Thanos, C. (eds.) *ECDL 2002*. LNCS, vol. 2458, pp. 292–308. Springer, Heidelberg (2002)

6. Dunn, J.W., Mayer, C.A.: Variations: a digital music library system at indiana university. In: DL 1999: Proceedings of the Fourth ACM Conference on Digital Libraries, pp. 12–19. ACM, New York (1999)
7. Agosti, M., Berretti, S., Brettlecker, G., Bimbo, A.D., Ferro, N., Fuhr, N., Keim, D.A., Klas, C.P., Lidy, T., Milano, D., Norrie, M.C., Ranaldi, P., Rauber, A., Schek, H.J., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: Delosdlms - the integrated delos digital library management system. In: DELOS Conference, pp. 36–45 (2007)
8. Berndt, R., Blmel, I., Krottmaier, H., Wessel, R., Schreck, T.: Demonstration of user interfaces for querying in 3d architectural content in PROBADO3D. In: 13th European Conference on Digital Libraries (2009)(Demonstration Paper)
9. PsychData National Repository for Psychological Research Data (in German), <http://psychdata.zpid.de/>
10. PANGAEA Publishing Network for Geoscientific & Environmental Data, <http://www.pangaea.de/>
11. Dryad Digital Repository for Data Underlying Published Works, <http://www.datadryad.org/>
12. Brase, J.: Using digital library techniques-Registration of scientific primary data. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 488–494. Springer, Heidelberg (2004)
13. ELIXIR European Life Sciences Infrastructure for Biological Information, <http://www.elixir-europe.org/>
14. Bamboo Research Initiative, <http://projectbamboo.org/>
15. Liao, T.W.: Clustering of time series data-a survey. *Pattern Recognition* 38, 1857–1874 (2005)
16. Agrawal, R., Lin, K., Sawhney, H., Shim, K.: Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In: Proceedings of the International Conference on Very Large Data Bases, Citeseer, pp. 490–501 (1995)
17. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Bertino, E., Christodoulakis, S., Plexousakis, D., Christophides, V., Koubarakis, M., Böhm, K., Ferrari, E. (eds.) EDBT 2004. LNCS, vol. 2992, pp. 676–693. Springer, Heidelberg (2004)
18. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2003)
19. Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization* 3(1), 1–18 (2004)
20. Schreck, T., Bernard, J., Von Landesberger, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization* 8(1), 14–29 (2009)
21. Kohonen, T.: *Self-Organizing Maps*, 3rd edn. Springer, Heidelberg (2001)
22. Šimunić, K.: Visualization of stock market charts. In: Proc. Int. Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (2003)
23. World Data Center System, <http://www.ngdc.noaa.gov/wdc/>
24. Baseline Surface Radiation Network (BSRN), <http://www.bsrn.awi.de/>
25. Ben, S.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proc. of the 1996 IEEE Symposium on Visual Languages, pp. 336–343. IEEE Computer Society, Washington (1996)
26. PANGAEA PanPlot Tool, <http://doi.pangaea.de/10.1594/PANGAEA.330147>

DINAH, A Philological Platform for the Construction of Multi-structured Documents

Pierre-Édouard Portier and Sylvie Calabretto

Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
{pierre-edouard.portier,sylvie.calabretto}@insa-lyon.fr

Abstract. We consider how the construction of multi-structured documents implies the definition of structuration vocabularies. In a multi-users context, the growth of these vocabularies has to be controlled. Therefore, we propose using the trace of users activity to limit this growth and document the vocabularies. A user will, for example, be able to follow and annotate the track of a vocabulary concept: from its creation to the last time it was used. From a broader point of view, this work is grounded on our Web based philological platform, DINAH, and is mainly motivated by our collaboration with a group of philosophers studying the handwritten manuscripts of Jean-Toussaint Desanti.

1 Introduction

We study how multi-structured documents are constructed in a multi-users context composed of philologists. Our work is based on experience gained working with philosophers who are building a digital edition of the handwritten archives of French philosopher Jean-Toussaint Desanti (1914-2002). Digital editing covers the whole editorial, scientific and critical process that leads to the publication of an electronic resource. In the case of manuscripts, editing mainly consists in the transcription and critical analysis of digital facsimiles, that is to say the creation of a textual document associated with the images of a handwritten manuscript. We found that the problem of constructing multi-structured documents was at the heart of their work. Indeed, they need to let coexist a multiplicity of structures in order to be able to access a document according to many interpretations. First, we will describe a methodology that promotes the emergence of multiple structures in a multi-users context. Then, we will introduce a dynamic documentation mechanism that can be used to control the growth of structuration vocabularies.

2 Construction of Multi-structured Documents

We define the notion of multi-structured documents and describe the problem of their representation. Then, we introduce a methodology for their construction.

2.1 Multi-structured Documents

Definitions

A *resource* is anything uniquely identified by an URI. Fragments, intervals, zones, terms, classes, binary relations, structuration vocabularies and documents are resources.

A *fragment* is a part of document content. Our documents are textual documents and manuscripts images. In the case of textual documents a fragment is the pair $(D, (inf, sup))$ where D is a document identifier, and (inf, sup) is an *integer interval* addressing a part of the document. In the case of images a fragment is the pair $(I, ((x1, y1), (x2, y2)))$ where I is an image identifier and $((x1, y1), (x2, y2))$ are the coordinates of a *rectangular zone* of the image.

A *term* is a string of characters and a *class* is a set of terms. A *binary relation* $R(x, y)$ links together two resources and a *structuration vocabulary* is a set of binary relations. Finally, a *multi-structured document* is a document with fragments participating in relations that belong to multiple structuration vocabularies.

Before proceeding further, we should exemplify the previous definitions. It is also the occasion to introduce some functionalities of our philological software platform named DINAH. Consider the following scenario: a philologist finds a consistent subset about Marx inside a stack of pages of consequent size. He isolates this subset by creating a new collection (see figure 1). He creates a relation "mainSubject" between this collection and the term "marx" from the class "Author". He begins to transcribe the collection and also creates relations, such as "quotation", "citationTitle", between intervals of the transcribed text and the document (see figure 2). He discovers later that this collection is in fact a preparation for another work he found in the archive. He creates a relation "preparationFor" between the two collections (see figure 3). Etc. Etc. These newly created relations dynamically update the faceted navigation interface that can be used to find specific collections or pages by iterative refinement (see figure 4).

How is it that, for example, a user chooses to place the relation "citationTitle" within the "citations" vocabulary while he affects the relation "hasLine" to the "physicalStructure" vocabulary? In a multi-users context, how a user will know the meaning of a relation created by someone else? We will address the first question in the remaining parts of this section, and the second question in the next section. We should now recall some characteristics of the existing models for the representation of multi-structured documents.

Existing Models

Multi-structured documents have to be analyzed in their historical context where the most used formalisms for documents representation (first SGML then XML) implied tree structures. That is why this problem has so far been considered under the technical point of view of overlapping hierarchies. From our previous

example, let say a page has been transcribed and relations have been created to indicate where citations occur. Then, the lines of text are isolated in order to align the transcription with the manuscript facsimile. It might happen that a quotation overlaps two lines and there would be locally a graph structure: a natural use of XML becomes impossible (see figure 5). We now describe different solutions for the representation of multi-structured documents.



Fig. 1. Creation, reordering, navigation and annotation of collections of images. Subject (or object) of the relation is dragged on the subject (or object) label, the relation itself is chosen (or created if it didn't exist) from an autocomplete menu.

We divide the set of existing solutions into four classes: historical solutions, ad-hoc solutions, models not compatible with XML and finally models compatible with XML. We characterize each solution according to four criteria. The first one is the "genericity" and determines, when a model exists, if we can modify it in order to manage problems outside of the initial scope of multi-structured documents representation. The second criterion measures the quality of the implementation of the solution. The third is about the existence and effectiveness of "query mechanisms" for multi-structured documents. The last criterion determines if the model is robust to change of document content or document structures.

CONCUR [1] is a feature of SGML designed to allow the integration inside a same document of tags extracted from different DTDs. Thus, if the definitions of the overlapping tags appear in different DTDs, the representation problem of multi-structured documents is solved. However, because of its complexity, this SGML proposal has never been entirely implemented.

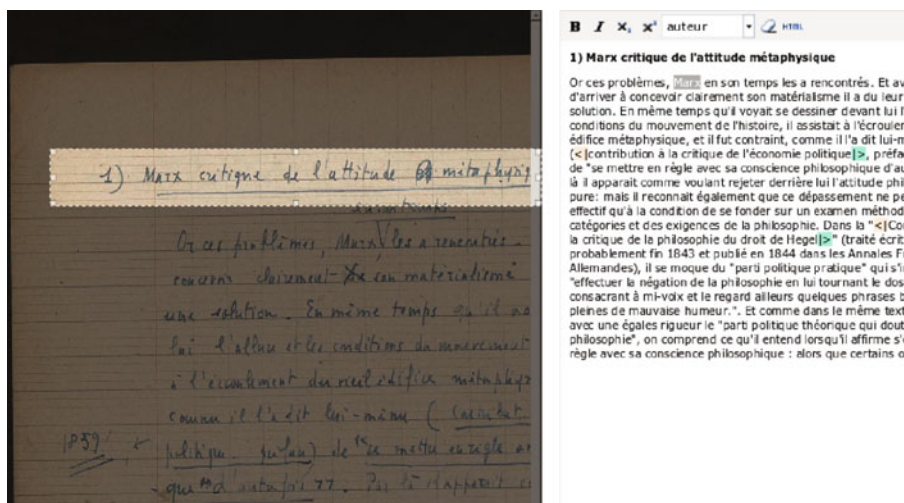


Fig. 2. Transcription and annotation of a manuscript page

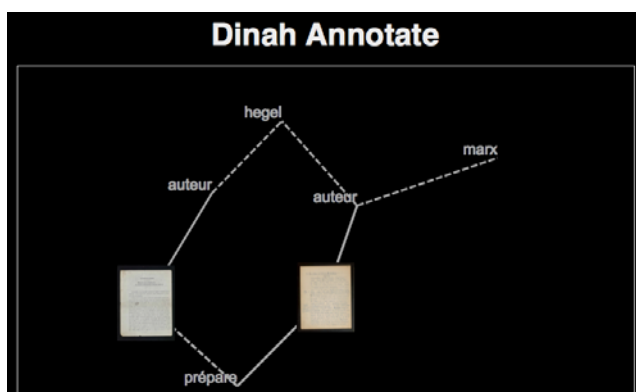


Fig. 3. Visualization of relations

The TEI [\[2\]](http://www.tei-c.org/)¹ describes different syntactic solutions for the representation of multiple hierarchies into the same text (as the use of milestones or the fragmentation of elements, etc.). The main disadvantage of this solution is the impossibility to effectively use the standard XML tools (XQuery, XPath, ...) with the resulting multi-structured documents.

Since the main problem for the representation of multi-structured documents seems to be the syntactic limitations of XML, some solutions are based on models with alternative syntaxes. However they cannot profit from the galaxy of

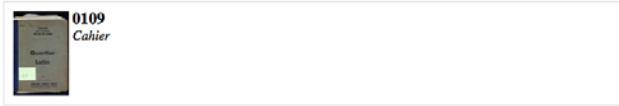
¹ Text Encoding Initiative. <http://www.tei-c.org/>

Les collections J-T. Desanti

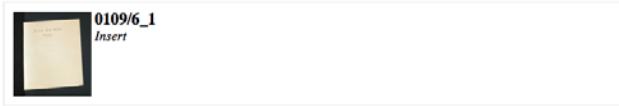
1019 Items

Trier par : libellés, puis par... • ☑ Grouper selon le tri

0109 (1)



0109/6_1 (1)



0109/6_1/2_1 (1)



Type

- 91 Cahier
- 9 Collection Perso
- 724 Insert
- 195 Pochette

auteur

1011 (missing this field)

- 1 archimède
- 3 aristote
- 1 augustin
- 1 buchelard
- 1 benveniste
- 2 bergson

concept

1017 (missing this field)

- 1 finalité
- 1 formalisme
- 1 histoire
- 1 langage
- 1 ...

Fig. 4. Navigation among collections

<line>exigences de la philosophie. Dans la *Contribution à la critique* de la philosophie du droit de Hegel, il se moque ...

Fig. 5. Illustration of overlapping hierarchies

tools offered by XML. Among those solutions, we can distinguish LMNL [3] and TexMecs [4] which are alternatives to XML (formal models and syntaxes) specifically designed for the representation of overlapping structures, from propositions that take advantage of the native graph model of RDF to represent multi-structured documents. Among these, the most convincing certainly is EARMARK [5]. The notions of "location", "range", "markup item", etc. used for modelling multi-structured documents are precisely defined in an OWL ontology. Moreover, the SPARQL language can be used to query the documents. It is to be noted that the origins of the EARMARK proposal are to be found in

two previous works: annotations graphs [6] are used, in the context of linguistic research, to represent documents as graphs so as to avoid the overlapping hierarchies problem ; RDFTef [7] can be seen as an adaptation of annotations graphs for the RDF standard formalism.

Finally there are solutions that remain compatible with XML but either extend the XML model itself or modify some XML tools (such as XPath and XQuery) to work with multi-structured documents. Representatives of the first category, the multi-colored trees [8] and the delay nodes [9] solutions have very similar models based on an extension of the core XML model to consider documents as set of XML trees. But unlike multi-colored trees, for delay nodes no XPath extension is necessary in order to navigate inside the structures. We now introduce members of the second category (documents syntactically expressed with XML but accompanied by modified XML tools to operate on them). GODDAG [10] (General Ordered Descendant Directed Acyclic Graph), MSXD [11], MonetDB [12] and MultiX [13] are similar proposals since in each case several trees are defined over the same textual content by sharing their leaves (textual fragments). MSXD introduces for the first time the idea of a schema for multi-structured documents. The MonetDB proposal is an extension to the MonetDB/XQuery XML SGBD with optimized query operators added to XPath with four new axis steps. These steps have been implemented very efficiently by using a region index and fast algorithms. MSDM is a lightweight solution that needs no more than a few specialised XQuery functions. Each one of these four previous solutions fails at managing change in content or structures since the entire structures have to be reconstructed every time modifications happen. MuLaX [14] is an adaptation of the previously described SGML CONCUR option to the XML world. An editor has been developed as an Eclipse plugin for the creation of MuLaX documents, but no query mechanism has been defined. Finally, feature structures [15] are a general purpose knowledge representation format that can be used as a representation format for XML documents annotated with heterogeneous tag sets, it was adopted as a standard by the TEI in 2006. Feature structures have solid mathematical foundations. In particular the two operations of unification and generalisation are well defined and offer very interesting perspectives for the combination of multi-structured documents. However, there is no specialised query mechanism and no way of managing change in content or structures.

Table 1 summarizes the analysis by affecting, as objectively as possible, a score from 0 to 3 to each criterion (genericity, quality of implementation, query mechanisms, management of changes in data and structures), for each solution.

2.2 A Strategy for the Construction of Multi-structured Documents

The previous solutions help us understand what multi-structured documents are and how they can be represented, but none of them seem to be interested in the way structures appear! They must appear in the process of document construction. In a previous work [16] we designed a methodology for the creation and maintenance of multi-structured documents. It was based on a set of Haskell

Table 1. Rating of existing solutions for the representation of multi-structured documents

model		generi- city	implem- entation	query mecha- nisms	structure and data changes
TEI Guide- lines	redundant encod- ing	0	1	0	0
	empty elements	0	1	0	0
	virtual elements	0	1	0	0
CONCUR		0	1	0	2
MuLaX		0	2	1	2
TexMECS		0	2	1	2
LMNL		0	2	0	2
Delay Nodes		1	2	2	0
Annotations Graphs		2	2	2	2
RDF (RDFTEF)		3	1	1	2
EARMARK		3	1	3	3
MonetDB		1	3	3	1
MCT		2	2	2	1
Features Structures		3	1	1	1
MSXD		2	2	3	0
GODDAG		3	2	3	2
MSDM/MultiX		3	2	3	2

(a functional programming language) functions. Since then, significant changes occurred. We will explain on the previous example of a multi-structured document (see figure 5) how we now model this process of document construction. First of all, we have to say that from the previous analysis we choose to represent our documents in the RDF formalism but, as it will be understood in the following explanation, we voluntarily impose each structure to be hierarchical (as for the MultiX, MSXD and GODDAG solutions).

We saw that the technical issue of multi-structured documents is the one of overlapping hierarchies. Moreover, if we do not consider the documents as immutable objects but as dynamic objects that have to be constructed, we must admit the fact that overlapping hierarchies must happen at precise times. We should take an example. Let say a user annotated some citations titles and quotations he found in his transcription of a manuscript. Later he is told that in order to precisely align his transcription with the original facsimile he should annotate each line of the manuscript. So, he begins this new annotation task and since the “line” relation did not exist he adds it to the current vocabulary (the one already containing “citationTitle”, “quotation”, etc.). At some time, while he has already marked some lines, a new line he would like to describe overlaps with an existing citation title. Our system (DINAH) will then alert him about an incompatibility between the relations “citationTitle” and “line” and advice him to assign either “citationTitle” or “line” to another, and possibly new, vocabulary. In this case, he may assign “line” to a “physical structure” vocabulary. Figure 6 is a sample of the resulting graph.

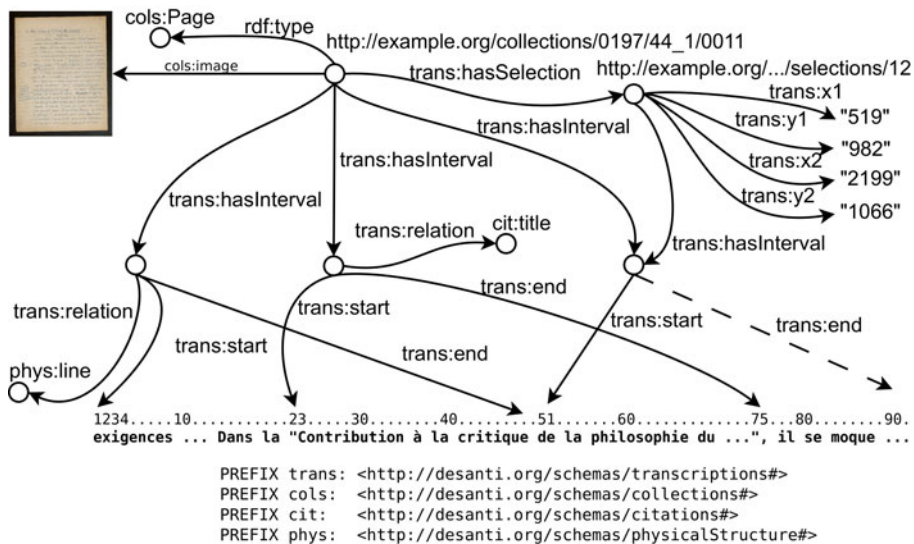


Fig. 6. Sample from our RDF representation of multi-structured documents

Finally, our strategy for the management of multi-structured documents promotes the construction of a multiplicity of structures that should reflect the perspectives adopted by the users while accessing the documents. Each user has the liberty to create new vocabularies. Moreover, when overlapping hierarchies are detected they are encouraged to solve the problem by introducing a new vocabulary. In our multi-users context, this liberty could lead to an uncontrolled growth of vocabularies with lots of duplicate usages, synonyms, etc. That is why the next section present a proposal for the dynamic documentation of structuration vocabularies.

3 Reflexions on Structuration Vocabularies

3.1 Dynamic Documentation

Our idea for the dynamic documentation of structuration vocabularies relies on the monitoring of user actions. When a user wants to know how to use a term or a relation he can ask for a representation of the trace of users actions centered on the action that leads to the term (or relation) creation or any instances of its use. This trace can itself be annotated. Users benefit from this last kind of annotations to document the vocabularies (see figure 7). Most of the time the user who document a term or a relation is the one who first created it. In case of multiple annotations they are ordered by the name of the annotator.

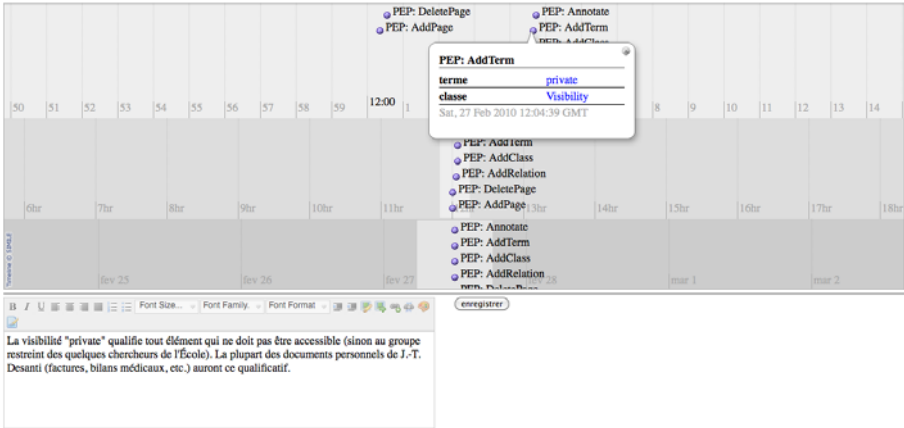


Fig. 7. Visualization of the trace of user activities

3.2 Trace Model

Existing approach. There are few works dealing with the use of activity traces for knowledge management ([17] being one of the most representative). They insist on the reflexive nature of the "use traces" as a way to share knowledge. They also define generic (and quite complex) activities models and transformations rules to go from the original trace to one with the right granularity level in order to be meaningful to the user. However, we choose to adopt a more lightweight approach well adapted to our needs.

A lightweight model. We define a simple RDF vocabulary to represent actions (see listing [1.1] in the turtle RDF syntax). The only requirement is that each time a developer add a new Action to the system he has to create sub-properties of the "withArgument" property for each argument of the new action. We then use simple SPARQL queries to build representations of the trace (see figure [7]).

Listing 1.1. Our trace model

```

PREFIX users: <http://desanti.org/schemas/users#>
PREFIX traces: <http://desanti.org/schemas/traces#>
INSERT INTO <http://desanti.org/> {
  traces:Action          a          rdfs:Class .
  traces:hasDoer         a          rdf:Property .
  traces:hasDoer         rdfs:domain traces:Action .
  traces:hasDoer         rdfs:range  users:User .
  traces:hasTimestamp   a          rdf:Property .
  traces:hasTimestamp   rdfs:domain traces:Action .
  traces:withArgument   a          rdf:Property .
  traces:withArgument   rdfs:domain traces:Action .
  traces:documentation  a          rdf:Property .
  traces:documentation  rdfs:domain traces:Action .
  traces:documentation  rdfs:range  rdfs:Literal .
  traces:withInterval   rdfs:subPropertyOf traces:withArgument .
  traces:withInterval   rdfs:range  trans:Interval .
  traces:withInterval   rdfs:label  "intervalle" .
}
    
```


4 Comparison with Existing Philological Platforms

Though this work deals mainly with the creation of multi-structured documents, it remains generic enough and can be compared to other philological platforms. We divide them in two categories: first platforms of historical interest, next Web based platforms.

4.1 Historical Platforms

BAMBI [18] (Better Access to Manuscripts and Browsing of Images) is, according to the authors, "an hypermedia system allowing historians to read and transcribe manuscripts, write annotations, and navigate between the words of the transcription and the matching piece of image in the facsimile of the manuscript". It was the first philological software platform. It does not allow typed annotations.

Part of the DEBORA [19] (Digital Access to Books of the Renaissance) project consisted in a digital library system with collaborative features. It introduced the notion of "virtual books". A virtual book is the representation of a path among pages of the entire archive. But they are not resources themselves, they cannot be annotated. However we can consider this system as a first step towards a reflexive system that places users in front of their own activities.

HyperNietzsche [20] (today named Nietzschesource) was a pioneer digital library platform. A path mechanism is present, very similar to the virtual books of the DEBORA project. However as for the virtual books, the paths are not resources and thus cannot truly enter in a collaborative process that would allow to exchange and annotate them.

4.2 Web Based Platforms

Collate [21], TALIA [22], PINAKES [23], BRICKS [24] and JeromeDL [25] are philological platforms based on semantic Web technologies. They offer high quality mechanisms for collaborative annotations. But they do not provide convergence mechanisms to isolate and document annotations vocabularies.

Armarius [26] is used to classify and annotate collections of manuscripts. It only provides untyped generic annotations. But it offers a view of all the user actions that occurred during the current session and plans to apply graph matching algorithms in order to, for example, deduce probabilities for the next actions. Thus, it can be compared with our use of traces.

5 Conclusions

We introduced the little-studied problem of multi-structured documents construction. We did not follow the conventional view that considers the heart of the problem to be the technical difficulty of representing overlapping hierarchies. On the contrary, we chose to consider overlapping hierarchies events as triggers for the creation of new structures. Furthermore, in order to manage the growth of structuration vocabularies we introduced a dynamic documentation mechanism based on the users traces of actions. Finally, all the propositions have been implemented in our philological software platform named DINAH.

References

1. Goldfarb, C.F.: The SGML handbook. Oxford University Press, Inc., New York (1990)
2. Burnard, L., Bauman, S.: Tei p5: Guidelines for electronic text encoding and interchange (2007)
3. Tension, J., Piez, W.: The layered markup and annotation language (lxml). In: Extreme Markup Languages (2002)
4. Huitfeldt, C., Sperberg-McQueen, M.: Texmecs: An experimental markup meta-language for complex documents (2003)
5. Peroni, S., Vitali, F.: Annotations with earmark for arbitrary, overlapping and out-of order markup. In: Borghoff, U.M., Chidlovskii, B. (eds.) ACM Symposium on Document Engineering, pp. 171–180. ACM, New York (2009)
6. Maeda, K., Bird, S., Ma, X., Lee, H.: Creating annotation tools with the annotation graph toolkit. In: Proceedings of the Third International Conference on Language Resources and Evaluation (April 2002)
7. Tummarello, G., Morbidoni, C., Pierazzo, E.: Toward textual encoding based on rdf. In: ELPUB, pp. 57–63 (2005)
8. Jagadish, H.V., et al.: Colorful xml: one hierarchy isn't enough. In: SIGMOD 2004: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 251–262. ACM, New York (2004)
9. Le Maitre, J.: Describing multistructured xml documents by means of delay nodes. In: DocEng. 2006: Proceedings of the 2006 ACM Symposium on Document Engineering, pp. 155–164. ACM, New York (2006)
10. Sperberg-McQueen, C.M., Huitfeldt, C.: Goddag: A data structure for overlapping hierarchies. In: DDEP/PODDP, pp. 139–160 (2000)
11. Bruno, E., Murisasco, E.: Multistructured xml textual documents. GESTS International Transactions on Computer Science and Engineering 34(1), 200–211 (2006)
12. Alink, W., Bhoedjang, R.A.F., de Vries, A.P., Boncz, P.A.: Efficient xquery support for stand-off annotation. In: XIME-P (2006)
13. Chatti, N., Kaouk, S., Calabretto, S., Pinon, J.M.: MultiX: an XML-based formalism to encode multi-structured documents. In: Proceedings of Extreme Markup Languages 2007, Montréal, Canada (August 2007)
14. Hilbert, M., Witt, A., Québec, M., Schonefeld, O.: Making concur work. In: Extreme Markup Languages (2005)
15. Stegmann, J., Witt, A.: Tei feature structures as a representation format for multiple annotation and generic xml documents. In: Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (August 2009), doi:10.4242/BalisageVol3.Stegmann01
16. Portier, P.E., Calabretto, S.: Creation and maintenance of multi-structured documents. In: DocEng. 2009: Proceedings of the 9th ACM Symposium on Document Engineering, pp. 181–184. ACM, New York (2009)
17. Laffaquière, J., Settouti, L.S., Prié, Y., Mille, A.: Trace-based framework for experience management and engineering. In: KES (1), pp. 1171–1178 (2006)
18. Bozzi, A., Calabretto, S.: The digital library and computational philology: The bambi project. In: Peters, C., Thanos, C. (eds.) ECDL 1997. LNCS, vol. 1324, pp. 269–285. Springer, Heidelberg (1997)
19. Nichols, D.M., et al.: Debora: developing an interface to support collaboration in a digital library. In: Agosti, M., et al. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 239–248. Springer, Heidelberg (2009)

20. D'Iorio, P.: Nietzsche on new paths: The hypernietzsche project and open scholarship on the web. In: Fornari, M.C., Franzese, S. (eds.) *Friedrich Nietzsche. Edizioni e interpretazioni*, Pisa ETS (2007)
21. Stein, A., Keiper, J., Bezerra, L., Brocks, H., Thiel, U.: Collaborative research and documentation of european film history: The collate collaboratory. *International Journal of Digital Information Management (JDIM)*, special issue on aWeb-based collaboratories from centres without, 30–39 (2004)
22. Hahn, D., Nucci, M., Barbera, M.: The talia library platform - rapidly building a digital library on rails. In: *4th Workshop on Scripting for the Semantic Web* (2008)
23. Scotti, A., Nuzzo, D.: Pinakes – a modeling environment for scientific heritage database applications. In: *Proc. of Reconstructing Science – Contributions to the Enhancement of the European Scientific Heritage Workshop*, Ravenna, Italy (2001)
24. Bertoncini, M.: On the move towards the european digital library: Bricks, tel, michael and delos converging experiences. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) *ECDL 2007. LNCS*, vol. 4675, pp. 440–441. Springer, Heidelberg (2007)
25. Kruk, S.R., Woroniecki, T., Gzella, A., Dabrowski, M.: Jeromedl - a semantic digital library. In: Golbeck, J., Mika, P. (eds.) *Semantic Web Challenge. CEUR Workshop Proceedings*, vol. 295, CEUR-WS.org (2007)
26. Doumat, R., et al.: Online ancient documents: Armarius. In: *ACM DocEng. 2008. Proceeding of the Eighth ACM Symposium on Document Engineering*, pp. 127–130. ACM, New York (September 2008)

The PROBADO Project - Approach and Lessons Learned in Building a Digital Library System for Heterogeneous Non-textual Documents

René Berndt¹, Ina Blümel², Michael Clausen³, David Damm³, Jürgen Diet⁴, Dieter Fellner^{5,6}, Christian Fremerey³, Reinhard Klein³, Frank Krahl⁴, Maximilian Scherer⁵, Tobias Schreck⁵, Irina Sens², Verena Thomas³, and Raoul Wessel³

¹ Graz University of Technology, Austria

² German National Library of Science and Technology Hannover, Germany

³ University of Bonn, Germany

⁴ Bavarian State Library, Munich, Germany

⁵ Technische Universität Darmstadt, Germany

⁶ Fraunhofer Institute for Computer Graphics, Darmstadt, Germany

Abstract. The PROBADO project is a research effort to develop and operate advanced Digital Library support for non-textual documents. The main goal is to contribute to all parts of the Digital Library work flow from content acquisition over indexing to search and presentation. While not limited in terms of supported document types, reference support is developed for classical digital music and 3D architectural models. In this paper, we review the overall goals, approaches taken, and lessons learned so far in a highly integrated effort of university researchers and library experts. We address the problem of technology transfer, aspects of repository compilation, and the problem of inter-domain retrieval. The experiences are relevant for other project efforts in the non-textual Digital Library domain.

1 Introduction

Digital Library technology offers many effective ways to handle document content. Access and delivery of documents becomes more and more digital and decentralized, and new user groups can benefit from library services. This is true for textual documents. However, technological and scientific progress contribute to increasing availability of non-textual documents, which are worthy of library-oriented treatment. Examples include digitization efforts in Cultural Heritage, production of scientific film, recording of orchestral performances, as well as masses of primary research data produced in the natural sciences. All of these non-textual documents, while being potentially relevant for library-oriented service, are more difficult to accommodate in a Digital Library system than their textual counterparts. Main challenges in supporting non-textual documents include questions of document representation, indexing and content-based accessing, and document presentation. Specifically, content-based access in non-textual

documents is a difficult problem as appropriate methods usually are application dependent and nontrivial to implement.

From the field of multimedia databases and multimedia visualization, many promising approaches have been proposed. But even if relevant document domains, use cases, and accommodation strategies have been identified, the problem of deploying such approaches within the operational context of a library operator needs to be solved. PROBADO aims at designing, developing and deploying Digital Library functionality for non-textual documents for a selection of use cases. At the same time, the project aims to propose a general reference architecture and protocol for consolidation of distributed non-textual document repositories of heterogeneous document types.

In this paper, we report on the approach taken and the experiences made during the first three and a half years of the PROBADO project. We systematically discuss the challenges that arose so far during the project, and sketch our solutions for them. The contribution of this paper is to offer a joint conceptual and practical perspective on a substantial Digital Library research and deployment effort.

2 Related Work

We briefly recall related work on Digital Library systems and Multimedia Retrieval. Additional related work specific to the domains discussed throughout this paper is recalled in the corresponding paper sections.

Existing Digital Library systems include Fedora [12], Greenstone [20], DLib [4] and Variations [8]. Fedora, Greenstone, and DLib support building Digital Libraries for textual documents; support for multimedia documents relies on metadata annotations according to specific standards such as MPEG-7. In PROBADO, the goal is to index and access non-textual documents specifically by *content-based* approaches. Therefore, the aforementioned systems are not directly applicable to our approach.

In multimedia retrieval, commercial systems and research prototypes exist. Examples include Google's *Similar Images* and *3D Warehouse*, both of which allow for content-based search. VICTORY [6] is a research project developing content-based retrieval of 3D data using a peer-to-peer architecture. Multimedia retrieval systems such as these employ the same basic approach as PROBADO for supporting content-based search. Given a multimedia query (e.g. an example document), the system computes a mathematically tractable representation (descriptor) for this query and compares this to a database of descriptors of the indexed content. Details for search approaches in 3D and music retrieval as used in PROBADO are given in Sections 3.2 and 3.3.

3 The PROBADO Approach

PROBADO is a distributed multimedia Digital Library system developed jointly by university researchers and scientific library experts. PROBADO supports

metadata-based and content-based retrieval of *3D architectural models* and *classical music*. We give a concise review of the system components and the development and technology transfer approach.

3.1 Overview of the PROBADO System Architecture

The PROBADO framework is designed to integrate heterogeneous multimedia documents from distributed, specialized document repositories by means of a three layer architecture. User interface, middleware, and repository layers communicate by a SOAP-based web-service.

Users formulate content-based queries using document-dependent search interfaces provided by the repository layers. These queries are forwarded to the middleware. Any user interface needs to implement at least one of the search functions provided by the middleware. These query interfaces support either the search for textual metadata, the search for content-specific data or multi-modal search for both content and metadata [5]. The middleware layer forwards *content-based* queries to all connected repositories supporting the addressed search functions. *Metadata* queries are evaluated directly in the middleware, which hosts a consolidated index of metadata of all repositories. A synchronization mechanism keeps this metadata index up to date with the repositories. The repositories process the content-based queries. Result lists are returned to the middleware for aggregation and presentation to the user.

3.2 PROBADO 3D Repository

The PROBADO 3D Repository supports content-based indexing and retrieval in 3D architectural model data. It aims to support the architectural design process by searching in a Digital Library of architectural model data for re-usage, comparison and inspiration purposes. Useful content ranges from small furnishing objects to environmental elements up to building units and whole buildings.

Current approaches to 3D shape retrieval mainly focus on search for models that are geometrically similar to a query object. These methods are usually based on global or local shape descriptors. Additionally, view-based algorithms as well as graph-based approaches have been proposed. A detailed overview of state of the art methods in this area can be found in [15].

Data Preprocessing. During preprocessing, low-level technical metadata of the 3D model are extracted, previews are generated and for subsequent topological indexing, 3D building models are oriented and scaled consistently [3].

Content-based Indexing and Metadata. Content-based indexing allows searching in a query-by-example scenario and enables high-level metadata generation. For each model, a global shape descriptor is computed. Additionally, local shape descriptors are computed providing a high-quality object description, serving as a starting point for high-level metadata generation, eventually producing a *Room Connectivity Graph (RCG)* [19] characterizing their topology.



Fig. 1. (left) 2D result visualization. (right) Model details with integrated 3D preview.

From the RCG extraction phase, also high-level metadata like height of building models, the number of floors, doors, windows etc is obtained and stored for user access. Based on a supervised learning framework [18] using a preclassified 3D architecture benchmark [17], the model category is predicted and stored as well.

The 3D repository additionally stores metadata provided by the model creators including title, description, contributor information etc. These metadata together with the extracted semantic metadata can be queried for by means of simple and extended search forms.

Query-by-example. We currently provide four ways to formulate a query-by-example based on complete 3D models: (1) upload of example model; (2) a 3D sketch interface based on GML [2]; (3) a plug-in for the Google™ Sketchup modeling tool; and (4) using a previous query result as a query key. (2) is tailored to building models and based on searching the extracted RCGs for certain spatial arrangements of rooms and floors. We provide visual-interactive interfaces for all content-based search modalities as described in [1].

Result Visualization. Apart from traditional sequential result lists, the 3D layer currently provides a 2D visualization for results based on global object similarity, which is realized using multidimensional scaling. The details page for a selected result contains also a 3D preview based on PDF (see Fig. [1]).

3.3 PROBADO Music Repository

The PROBADO Music Repository supports content-based indexing and retrieval of digital classical music documents. This document notion includes different document types representing different aspects of a piece of music (e.g., sheet music, compact discs, and libretti). At the Bavarian State Library (Bayerische Staatsbibliothek, BSB) a digital collection of western classical music has been established. The collection currently contains approx. 96,000 pages of sheet music and corresponding audio recordings from compact disks. Facing such large multimodal digital document collections, systems to manage, process, browse, and

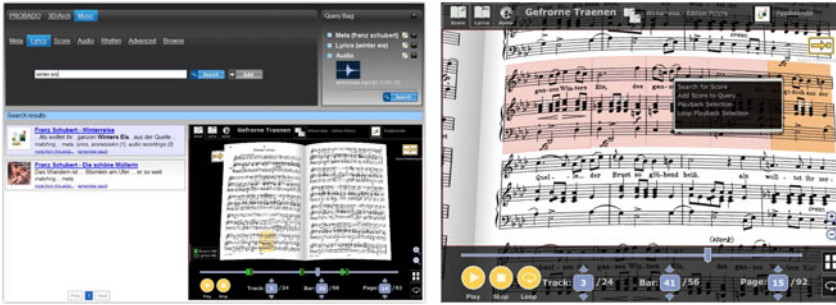


Fig. 2. (left) The PROBADO music frontend with integrated Score Audio Player. (right) The sheet music visualization can be used to perform content-based retrieval.

access this data are required. Within PROBADO, those requirements are being implemented. In addition, the well-established metadata search is expanded by offering content-based search functionalities.

Music information retrieval (MIR), amongst others, comprises the fields of content-based music retrieval and music alignment. The aim of content-based retrieval is to search for all occurrences of a query (e.g., melody, excerpt of a score, audio fragment) or slight variations thereof in a collection of music documents [10,14]. In the field of music alignment, different representations of the same piece of music are linked with each other, such that given a position within one document, the position within the other document describing the same musical position can be obtained [9,13,11]. For further literature on these and similar topics we refer to the proceedings of the annual ISMIR conference.

Applied MIR Techniques. In PROBADO we apply MIR techniques to preprocess a music document collection, to enable content-based retrieval, and to offer a holistic, attractive access to music documents. The developed preprocessing workflow provides a user interface for classical library tasks like metadata annotation. Moreover, automated MIR tasks are included (e.g., segmentation of scores, calculation of alignments between different music representations) [16].

Content-based Search Functionalities. For music documents, query engines are available, which process the following query formulations: (1) metadata; (2) lyrics; (3) audio fragments; (4) sheet music extracts; (5) a virtual piano to enter a music query.

Presentation. Presentation of music documents is realized by the Score Audio Player applet [5,16] (SAP, Figure 2). Its goal is an integrated presentation of all music documents representing the same piece of music. Due to the alignment information, synchronized playback of an audio recording while highlighting the corresponding bar (measure) within the sheet music is supported, allowing score based navigation. Also, the user can switch between different recordings while maintaining the musical position. Using the sheet music visualization, a number

of bars can be selected and directly be used as content-based query. The SAP also provides a detailed view on the matching regions within the piece of music.

4 Lessons Learned

4.1 System Architecture

The architecture of a distributed Digital Library system faces several challenges including metadata abstraction, relevance feedback, and inter-domain retrieval. To integrate heterogeneous document types a consolidated meta data abstraction is crucial. The trade-off between few but generic metadata fields and more, possibly specialized fields has to be regarded. In PROBADO a decision in favor of a compact DC-oriented metadata set was taken, securing extensibility to new domains. Specific metadata queries are still possible by directly querying in the individual repositories, but a joint metadata search over all repositories is evaluated using the unified DC scheme.

Relevance feedback (RF) techniques are important to support effective retrieval in multimedia data, but are difficult to apply in a heterogeneous and distributed environment. Results to be given feedback about may originate from different repositories. But since a given repository usually does not have information about the content of other repositories, it cannot solely apply the RF optimization mechanisms. Therefore RF-techniques are not employed within PROBADO.

Searching for multimedia data across domain boundaries is an open research question. To formulate a query-by-example, which is to be evaluated in combinations of domains, a compatible query syntax is necessary. E.g., for a combined content-based query in 3D models and 2D image data, a common syntax could be based on 2D images, as any 3D model can be projected to a 2D image. For other domain combinations like 3D model data and classical music, no such projection exists. Nonetheless, textual annotations and query-by-text can support inter-domain retrieval. Automatically generating semantically meaningful textual annotations from multimedia content is another research challenge. Inter-domain retrieval by textual queries is possible in PROBADO, restricted to manually obtained textual metadata.

4.2 Two Alternative Approaches to Repository Compilation

Our project includes the compilation of document repositories for each domain for three main purposes: (1) a reference collections for development and testing; (2) serve for demonstration purposes, raising interest; (3) obtain experience with digitizing and obtaining of documents from external providers.

The music reference collection is a large-scale digitization effort carried out in-house with the BSB library. For PROBADO purposes, this digitization workflow was augmented by an OMR-process ("optical music recognition"). The metadata model within the PROBADO music repository uses a work-centric data model that is based on the Functional Requirements of Bibliographic Records (FRBR) [7]. This *institution-oriented* approach is a highly structured process providing full control over the repository w.r.t. content, quality, and metadata.

The 3D repository comprises about 8,000 indexed models including buildings, construction units, furnishing etc. Providers include architectural component manufacturers, web portals for 3D content, and architecture faculties of universities. File formats, level of detail, content, quality, and the existence of metadata vary substantially. This *provider-oriented* approach is characterized by heterogeneity of the documents. Focus, format, resolution, and level of detail varies between documents.

5 Conclusions and Future Work

We reported on the approach and lessons learned in developing and deploying content-based Digital Library support for certain non-textual documents. While much has already been achieved in terms of functionality, selecting and transferring a suitable subset of functionality into practical operation represents organizational and technological challenges. Architectural and application implications relating to the distributed and heterogeneous system model, have been identified and were discussed. Two modes of repository compilation and two operation models were identified and compared.

Next steps involve actual transfer of functionality to the project library partners BSB and TIB, and customization of functionality for user needs. Steps in this stage include: (1) selection and consolidation of system functionality to be deployed, from the larger pool of developed functionality; (2) shaping the interfaces of the components to suit the hosting operational environment; (3) documentation and training of librarians and IT technicians; and (4) testing and usability iterations.

Due to the middleware abstraction layer, our approach does not restrict the supported document model. Consequentially, integration of additional document repositories is possible and will be aimed at. In the long run, research questions relating to the development of a document model supporting retrieval, presentation, and annotation of collections of heterogeneous non-textual documents need to be addressed.

Acknowledgments

PROBADO is a joint research project supported by the German Research Foundation DFG under the LIS program. PROBADO started in February 2006 with a tentative duration of five years. Sven Havemann, Harald Krottmaier, Frank Kurth, and Thorsten Steenweg made valuable contributions to the project effort. For further information, please visit the project website at <http://www.probado.de/>.

References

1. Berndt, R., Blümel, I., Krottmaier, H., Wessel, R., Schreck, T.: Demonstration of user interfaces for querying in 3d architectural content in PROBADO3D. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 491–492. Springer, Heidelberg (2009)

2. Berndt, R., Havemann, S., Fellner, D.: 3D Modeling in a Web Browser to Formulate Content-Based 3D Queries. In: Behr, J., Walczak, K. (eds.) *Proceeding of the 14th International Conference on 3D Web Technology*, Eurographics Association, Darmstadt (2009), <http://www.eg.org/EG/DL/PE/WEB3D09/111-118.pdf>
3. Berndt, R., Blümel, I., Wessel, R.: Probado3d towards an automatic multimedia indexing workflow for architectural 3d models. In: *14th International Conference on Electronic Publishing*, Helsinki (June 2010)
4. Castelli, D., Pagano, P.: Opendlib: A dl service system. In: Agosti, M., Thanos, C. (eds.) *ECDL 2002. LNCS, vol. 2458*, p. 292. Springer, Heidelberg (2002)
5. Damm, D., Kurth, F., Fremerey, C., Clausen, M.: A concept for using combined multimodal queries in digital music libraries. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) *ECDL 2009. LNCS, vol. 5714*, pp. 261–272. Springer, Heidelberg (2009)
6. Daras, P., Tzovaras, D., Dobravec, S., Trnkoczy, J., Sanna, A., Paravati, G., Trapfoener, R., Franz, J., Kastrinogiannis, T., Malavazos, C., Ploskas, N., Gumz, M., Geramani, K., Wintterle, G.J.: Victory: a 3d search engine over p2p and wireless p2p networks. In: *4th International Conference on Wireless Internet* (2008)
7. Diet, J., Kurth, F.: The PROBADO music repository at the Bavarian State Library. In: *8th International Conference on Music Information Retrieval* (2007)
8. Dunn, J.W., Byrd, D., Notess, M., Scherle, R.: Variations2: Retrieving and using music in an academic setting. *Communications of the ACM* 49 (2006)
9. Hu, N., Dannenberg, R., Tzanetakis, G.: Polyphonic audio matching and alignment for music retrieval. In: *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA* (2003)
10. Kurth, F., Müller, M.: Efficient index-based audio matching. *IEEE Transactions on Audio, Speech, and Language Processing* 16, 382–395 (2008)
11. Kurth, F., Müller, M., Fremerey, C., Chang, Y., Clausen, M.: Automated synchronization of scanned sheet music with audio recordings. In: *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR* (2007)
12. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.* 6, 124–138 (2006)
13. Orio, N.: Alignment of performances with scores aimed at content-based music access and retrieval. In: Agosti, M., Thanos, C. (eds.) *ECDL 2002. LNCS, vol. 2458*, p. 479. Springer, Heidelberg (2002)
14. Suyoto, I., Uitdenboger, A., Scholer, F.: Searching musical audio using symbolic queries. *IEEE Transactions on Audio, Speech, and Language Processing* 16 (2008)
15. Tangelder, J.W., Veltkamp, R.C.: A survey of content based 3d shape retrieval methods. *Multimedia Tools and Applications* 39, 441–471 (2008)
16. Thomas, V., Fremerey, C., Damm, D., Clausen, M.: SLAVE: a Score-Lyrics-Audio-Video-Explorer. In: *Proceedings of the 10th ISMIR* (2009)
17. Wessel, R., Blümel, I., Klein, R.: A 3d shape benchmark for retrieval and automatic classification of architectural data. In: *EG Workshop on 3D Object Retrieval* (2009)
18. Wessel, R., Baranowski, R., Klein, R.: Learning distinctive local object characteristics for 3d shape retrieval. In: *Vision, Modeling, and Visualization* (2008)
19. Wessel, R., Blümel, I., Klein, R.: The room connectivity graph: Shape retrieval in the architectural domain. In: *WSCG* (2008)
20. Witten, I.H., Mcnab, R.J., Boddie, S.J., Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In: *Proceedings of the Fifth ACM International Conference on Digital Libraries* (2000)

Capacity-Constrained Query Formulation

Matthias Hagen and Benno Maria Stein

Faculty of Media
Bauhaus University Weimar, Germany

Abstract. Given a set of keyphrases, we analyze how Web queries with these phrases can be formed that, taken altogether, return a specified number of hits. The use case of this problem is a plagiarism detection system that searches the Web for potentially plagiarized passages in a given suspicious document. For the query formulation problem we develop a heuristic search strategy based on co-occurrence probabilities. Compared to the maximal termset strategy [3], which can be considered as the most sensible non-heuristic baseline, our expected savings are on average 50% when queries for 9 or 10 phrases are to be constructed.

1 Introduction

The problem considered in this paper appears as an important sub-task of automatic text plagiarism detection. Plagiarized passages in a suspicious document can be found via direct comparisons against potential source documents. Today's typical source of plagiarism is the Web, which obviously contains too many documents for direct comparisons. The straightforward solution is to extract keyphrases from the suspicious document and to retrieve a tractable number of documents containing these phrases. These documents are considered as the best potential sources of plagiarism since they probably cover similar topics. Our contribution is a strategy for finding a family of “promising” Web queries whose combined results will be used for direct comparisons. The paper in hand does not deal with the complete plagiarism detection task; its focus is on the Web query pre-computation step.

The number of source documents a detection system can consider for direct comparisons is constrained by some processing capacity k . If all the extracted keyphrases (usually about 10) from the suspicious document are submitted as one single Web query, probably too few documents are returned with respect to k . Similarly, queries containing only few of the extracted phrases are likely to yield a huge number of hits; from these only a fraction, typically the Web search engine's top-ranked results, could be processed by the detection system. We argue that the probability to find potential plagiarism sources becomes maximum if the combined result list length of the promising queries is in the order of magnitude of the processing capacity k . We term this argument *the-user-knows-better hypothesis* or, more formally, *user-over-ranking hypothesis*: the detection system as the “user” of the search engine simply processes all of the promising queries' combined results, this way avoiding any search engine ranking issues that cannot be influenced.

Under the user-over-ranking hypothesis the CAPACITY CONSTRAINED QUERY FORMULATION problem analyzed in this paper is defined as follows. Given is (1) a

set W of keyphrases, (2) a Web search engine's query interface, and (3) an upper bound k on the number of desired documents. The task is to find a family $\mathcal{Q} \subseteq 2^W$ of queries, together returning at most k documents and containing all the phrases of W , if possible. Obviously, a series of queries must be submitted to the search engine for finding \mathcal{Q} , and we focus on the following optimization problem from the detection system's perspective: What strategy minimizes the average number of submitted queries? Two previous papers analyze related query formulation problems: Shapiro and Taksa [4] suggest the rather simple open end query formulation approach, for generating queries that each return at most an upper bound number of hits. Unfortunately, it is straightforward to construct situations in which the approach fails although adequate queries exist. A more involved maximal termset query formulation method is proposed by Póssas et al. [3]; we use an adapted version as our baseline.

2 Basic Definitions and the Baseline Method

Any subset $Q \subseteq W$ can be submitted as a Web query, with the notion that phrases are included in quotation marks. An engine's reply contains an estimation l_Q for the total number of results matching the query. Our task is to find a simple family $\mathcal{Q} = \{Q_1, \dots, Q_m\}$; *simple* means that $Q_i \not\subseteq Q_j$ for any $i \neq j$. Altogether \mathcal{Q} 's queries should not yield more than k results. From k we will derive an upper bound l_{\max} with the notion that a single query Q is promising iff $l_Q \leq l_{\max}$. Another lower bound l_{\min} is introduced for convenience reasons. We say that for $l_Q < l_{\min}$ the query Q is *underflowing*, whereas for $l_Q > l_{\max}$ it is *overflowing*. Queries that are neither under- nor overflowing are *valid*. A valid query Q is *minimal* iff omitting any phrase will result in an overflowing query. We propose the family \mathcal{Q}_{10} of all the minimal valid queries as a solution to CAPACITY CONSTRAINED QUERY FORMULATION. \mathcal{Q}_{10} is simple and covers all phrases that are contained in any valid query. During the computation we count the overall number *cost* of submitted Web queries.

Table 1. Keyphrase-document-relationships for the example scenario

Keyphrase	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
w_1	•	•		•	•					•
w_2		•				•			•	
w_3	•	•	•	•	•	•	•		•	
w_4	•			•		•	•	•		•
w_5	•		•		•	•	•	•	•	

Consider the following example scenario: Given are 10 indexed documents d_1, \dots, d_{10} and the set $W = \{w_1, \dots, w_5\}$ with the keyphrase-document-relationships shown in Table 1. Note that, submitted as a query, the set W itself will not result in any hit. Figure 1 shows a part of the hypercube of the possible 2^5 queries; the valid queries for $l_{\min} = 3$ and $l_{\max} = 4$ are shown highlighted. The query $\{w_3, w_5\}$ is overflowing (six hits) whereas $\{w_1, w_5\}$ is underflowing (two hits). The family

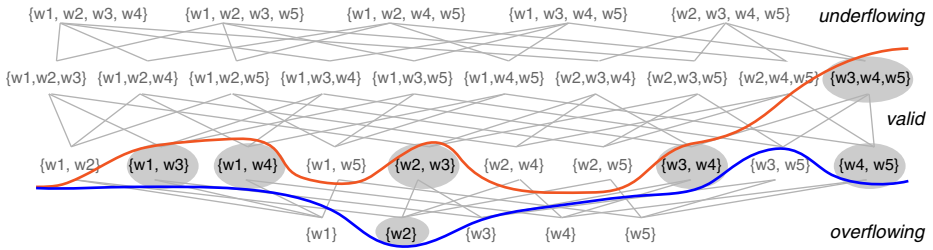


Fig. 1. Hypercube of possible queries in the example scenario

$Q_{l_0} = \{\{w_1, w_3\}, \{w_1, w_4\}, \{w_2\}, \{w_3, w_4\}, \{w_4, w_5\}\}$ corresponds to the lower border in Figure 1 it will return all documents except d_3 .

As a baseline we adapt the maximal termset approach by Póssas et al. [3], but we do not use GENMAX as a subroutine to enlarge promising keyphrase subsets. Instead, we adopt the classic Apriori algorithm that also stems from the field of frequent itemset mining [1] (cf. Figure 2 (left) for a basic pseudo-code listing). Apriori traverses the search space of all possible queries (cf. Figure 1) in a level-wise manner. Whenever the validity of a query Q has to be checked Apriori submits it to the Web search engine and obtains l_Q . The problem now is to find an appropriate l_{max} . We start with $l_{max} \leftarrow k$, compute Q_{l_0} using Apriori and count the results all queries in Q_{l_0} return. Usually this will be too many and we then use a binary search for an appropriate l_{max} by halving the value as long as the computed Q_{l_0} returns too many results. If at one intermediate step a Q_{l_0} returns approximately k results (we set the bound to at least 90%), the computation stops and outputs this Q_{l_0} . If eventually too few results are returned we enlarge l_{max} according to the binary search paradigm. Note that whenever we enlarge l_{max} for the first time, all of the remaining evaluations have already been done during the previous step such that no further queries have to be submitted.

Input: W, l_{min}, l_{max} **Output:** Q_{l_0}

if W is not overflowing **then**
 $Q \leftarrow \{\{w\} : w \in W \text{ and } \{w\} \text{ is valid}\}$
 $C_1 \leftarrow \{\{w\} : w \in W \text{ and } \{w\} \text{ is overflowing}\}$
 $i \leftarrow 1$
while $C_i \neq \emptyset$ **do**
 for all $Q, Q' \in C_i, |Q \cap Q'| = i - 1$ **do**
 $Q_{cand} \leftarrow Q \cup Q'$
 if $Q_{cand} \setminus \{w\} \in C_i$ for all $w \in Q_{cand}$ **then**
 if Q_{cand} is valid **then** $Q \leftarrow Q \cup \{Q_{cand}\}$
 if Q_{cand} overflows **then** $C_{i+1} \leftarrow C_{i+1} \cup \{Q_{cand}\}$
 $i \leftarrow i + 1$
output Q

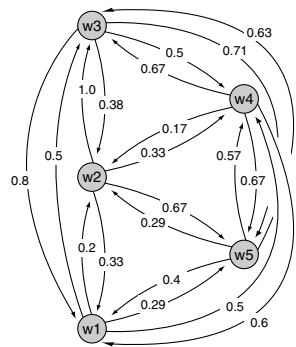


Fig. 2. Left: Apriori algorithm. Right: co-occurrence graph of Table 1's example scenario.

3 Outline of the Heuristic Search Strategy

To improve the performance of the baseline with respect to the number of submitted queries we propose a heuristic that mimics Apriori’s workflow but tries to avoid submission of Web queries. In a pre-processing step the heuristic derives a directed edge- and vertex-weighted co-occurrence graph G_W . The graph contains a vertex v_w for each keyphrase $w \in W$. The weight of v_w is set to $l_{\{w\}}$. An edge $e = v_w \rightarrow v_{w'}$ from v_w to $v_{w'}$ gets as weight the yield factor $\gamma(e) = l_{\{w,w'\}}/l_{\{w\}}$. Semantics: the yield factor multiplied by the weight of v_w gives the yield of Web hits when w' is added to the query $\{w\}$. Note that the yield factor is reminiscent of the co-occurrence probability for the keyphrases w and w' ; G_W is reminiscent of a mutual information graph (cf. Figure 2(right)). Obtaining G_W during pre-processing involves the same computations and Web queries that Apriori processes during the first two levels (queries with at most two keyphrases).

After the pre-processing step the heuristic starts an Apriori-like candidate generation on the third level (queries containing three phrases). Hence our technique does not save queries on the first two levels compared to Apriori, and no overall savings are achievable for initial keyphrase sets of size three. However, from Level 3 on G_W is used to assess a query before submitting it as a Web query. Assume we are on some level $i \geq 2$ (queries with i keyphrases). All processed queries Q from lower levels have a stored value est_Q indicating an estimation of the length of their result lists. Let the current candidate query Q_{cand} be obtained by merging queries Q and Q' from level $i - 1$. Before submitting Q_{cand} as a Web query (like the baseline would do) the estimation $est_{Q_{\text{cand}}} = est_Q \cdot \text{avg}\{\gamma(v_w \rightarrow v_{w'}) : w \in Q\}$ is computed. Submitting Q_{cand} as a Web query and obtaining the engine’s $l_{Q_{\text{cand}}}$ is done iff the estimation $est_{Q_{\text{cand}}}$ is in the order of l_{max} . Otherwise, no Web query is submitted; $est_{Q_{\text{cand}}}$ is remembered. This heuristic is not guaranteed to output the same family Q_{lo} as the baseline. However, experiments show good conformity of the output with the baseline’s Q_{lo} while saving a significant number of queries at the same time (see below).

4 Experimental Analysis and Conclusion

We experimentally compare our heuristic and the baseline as follows: for a given document we extract a number of keyphrases and then formulate queries using these phrases. Keyword extraction is managed by the head noun extractor [2]. Our document collection was obtained by crawling papers on computer science from major conferences and journals. We also added some books. From the established corpus we removed the documents for which we were not able to extract 10 reasonable keyphrases. Our test collection was formed by the 1112 remaining documents. We set the bounds $k = 1000$ and $l_{\text{min}} = 1$. For each document of the test collection we had 7 runs of the baseline and our heuristic with 4, 5, ..., 10 extracted keyphrases. Another run was done on the 775 documents of our collection from which 15 reasonable keyphrases could be extracted. As Web search engine we used the Microsoft Bing API. A typical Web query took about 300–550ms.

Table 2 shows the results of our experiments. For small keyphrase sets the complete query with all phrases often overflows (cf. first row). We filtered out such keyphrase

Table 2. Experimental results

Number of keyphrases:		4	5	6	7	8	9	10	15
1	Complete query overflows	647	535	444	351	274	229	212	18
2	Remaining documents	465	567	668	752	838	883	900	757
3	Avg. <i>cost</i> baseline	10.33	16.33	26.25	39.93	62.11	98.73	150.37	1 379.33
4	Avg. <i>cost</i> heuristic	11.09	19.66	36.33	64.92	117.05	207.20	342.95	3 020.34
5	Micro-averaged cost ratio	0.93	0.83	0.72	0.62	0.53	0.48	0.44	0.46

sets and derived the statistics (rows 3 to 5) for the remaining documents (number given in second row). In rows 3 and 4 we state the average number *cost* of Web queries the approaches submitted. The average ratio of submitted queries of the heuristic over the baseline is given in row 5. The possible savings are substantial: Even for 7 phrases our heuristic saves 30–40% of the queries and for 9 phrases possible savings reach 50%. Altogether, our results suggest that a near real-time plagiarism detection service with processing capacity $k = 1000$ should try to extract 9 or 10 keyphrases as then the heuristic computes \mathcal{Q}_{10} in about 1 minute.

References

- [1] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proc. of VLDB 1994, pp. 487–499 (1994)
- [2] Barker, K., Cornacchia, N.: Using noun phrase heads to extract document keyphrases. In: Proc. of AI 2000, pp. 40–52 (2000)
- [3] Póssas, B., Ziviani, N., Ribeiro-Neto, B.A., Meira Jr., W.: Maximal termsets as a query structuring mechanism. In: Proc. of CIKM 2005, pp. 287–288 (2005)
- [4] Shapiro, J., Taksa, I.: Constructing web search queries from the user’s information need expressed in a natural language. In: Matsui, M., Zuccherato, R.J. (eds.) SAC 2003. LNCS, vol. 3006, pp. 1157–1162. Springer, Heidelberg (2004)

AAT-Taiwan: Toward a Multilingual Access to Cultural Objects

Shu-Jiun Chen, Diane Wu, Pei-Wen Peng, and Yung-Ting Chang

Research Center for Information Technology Innovation,
Academia Sinica, Taipei 115, Taiwan
sophy@gate.sinica.edu.tw,
{taiyen,peiwen,yting}@iis.sinica.edu.tw

Abstract. This paper reports on current collaborative work between Taiwan e-Learning and Digital Archives Program (TELDAP) and Getty Research Institute (GRI) in developing the Chinese-language *Art & Architecture Thesaurus* (AAT-Taiwan) which supports the unification of terminology used by various archiving institutions for describing identical concepts. This work aims to establish a conceptual framework for the digital library by providing controlled vocabularies to index and catalogue the collection. With its multilingual nature, AAT Taiwan is able to bridge Western and Eastern culture in an integrated framework, and make our resources accessible worldwide. With its hierarchical structure, it also enhances the effectiveness and comprehensiveness of information retrieval in digital libraries.

Keywords: digital library, multilingual thesaurus, knowledge organization system.

1 Introduction

The idea of digital libraries prevailing nowadays prompts museums and organizations to digitize their collection and establish online databases for worldwide access. With millions of terms varied in domains and cultures, a single concept is frequently presented in various ways in terms of languages, which increases the hurdle to obtain the optimum result. The collaboration between the Taiwan e-Learning and Digital Archives Program¹ (TELDAP) and the Getty Research Institute², was therefore established in late 2008 to enhance the accessibility of this abundant collection. But first, the challenges of language barrier and integrating different terms used by various institutions and programs must be overcome. This paper illustrates how to utilize controlled vocabularies to establish a multilingual thesaurus that could widely disseminate the knowledge.

The researches on Thesauri as knowledge organization systems (KOS) are mostly about its multilingual interoperability in generic domains or between generic and

¹ Taiwan e-Learning and Digital Archives Program (TELDAP), <http://teldap.tw/en>

² The Getty Research Institute (GRI), <http://www.getty.edu/research/>

specific domains, such as Sinica Bow[1] and MACS[2]. Most KOS interoperability researches focus on the relationship between different languages in scientific domain, such as UMLS Metathesaurus[3] and AGROVOC[4], only a few KOS in Western languages are about arts and humanities, such as HEREIN[5]. Based on these research projects, multilingual KOS can be developed using methodologies of translation and mapping, with mapping methodology being the key to interoperability between different languages. By incorporating Eastern concepts and TELDAP collections into AAT-Taiwan³, the global Chinese-speaking communities would have access to a well-developed and authoritative information database.

2 Methods

The project methodology is derived from methods used in Linguistics, Semantics, Translation Studies, and Information Retrieval. It consists of five modules: Methodology R&D, Equivalence Mapping, Translation, Localization, Creation of New Concepts and Contribution, Technical and Functional development[6]. In Methodology R&D, workflows are formed to ensure optimum performance of each subsequent step. There are two approaches to select Chinese terms for Equivalence Mapping, from the controlled vocabularies used by institutions and organizations, or from authoritative references (Fig. 1, M1). Once the term is mapped to a matching Western concept (Fig. 1, M2), we then proceed to determine the equivalence mapping types (Fig. 1, M3) followed by Translation and Localization.

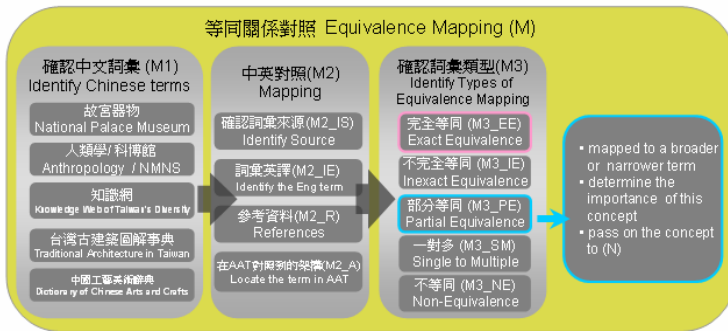


Fig. 1. The workflow of Equivalence Mapping

The Translation process includes English-to-Chinese translation, proofread, and expert check. Once a record is completed, it will be contributed to AAT⁴ via online contribution form or XML export for batch records. Each domain expert is required to provide at least two references for each of the following field: subject, prefer and

³ Art & Architecture thesaurus Taiwan, <http://aat.teldap.tw>

⁴ AAT, http://www.getty.edu/research/conducting_research/vocabularies/aat/

non-preferred terms, and scope note. This elaborate effort ensures every new record in the system is characterized by explicit specification toward building an ontology.

3 System Demo and Outcomes

The AAT-Taiwan system can export the XML format of a term with related data including preferred and non-preferred terms, broader and narrower terms, and related terms...etc. It is also compatible with SKOS data exchange standard.

This multilingual thesaurus contains American English, British English, Dutch, French, German, Spanish, and now Chinese. Every contribution record to the AAT includes 4 different language codes: Traditional Chinese, transliterated Wade-Giles, transliterated Hanyu Pinyin, and transliterated Pinyin. An AAT-Taiwan record display should contain a term (Fig.2-A), its equivalent to other languages (Fig 2-D), related images (Fig 2-B), scope note (Fig 2-C), hierarchical position(Fig 2-E), and external links (Fig 2-F). Each record also has a unique numeric identification that can be used to retrieve equivalent English term in the AAT[7].

The screenshot displays a detailed record for the concept 'standard script' (標準正體字) in the AAT-Taiwan system. The interface includes a header with navigation buttons, a main content area with various tabs, and a right-hand sidebar with detailed information. The main content area shows the term name, a table of related terms, and a form for adding new terms. The right-hand sidebar contains a scope note, a list of terms in different scripts, a hierarchical position tree, and external links.

Fig. 2. A complete record of an AAT-Taiwan concept

4 Discussion

As the project progresses, we aim to resolve the problems of insufficient Chinese references of Western-centered concepts, determining the importance of certain concepts in relation to our culture, and ensuring the accuracy of equivalence mapping results. The most noticeable problem with equivalence mapping is that every institution tends to implement its own classification methods and use of terms. More often than not, we need to re-examine and sometimes adopt a different set of logic in order to find matching concepts in AAT[8].

5 Conclusion and Future Work

Our goal is to provide global Asian cultural heritage communities with a knowledge-based database as reference to collection access and data value standard. In order to ensure the success of this international collaboration effort, a Multilingual Vocabulary Working Group has been set up in October 2009 by GRI as a means to establish close communication among international collaborators, including Centro de Documentación de Bienes Patrimoniales (DIBAM, Chile), Netherlands Institute for Art History (Netherlands), and State Museums of Berlin/Institute for Museum Research (Germany).

To bridge the lexical gap, AAT-Taiwan plans to collaborate with Chinese WordNet⁵ in the near future. With its concept driven and relation based lexical knowledge-base, AAT-Taiwan can develop a universal set of lexical semantic relations to execute multilingual query and disseminate information with same format regardless of different source languages. Another prospect is to implement information retrieval to organize matching data and trace back to the original sources. This technique can broaden the use of terminology in AAT-Taiwan and incorporate TELDAP collections, including metadata and digital images into its database[9].

References

1. Huang, C., Chang, R., Lee, S.: Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. In: Proceedings of the LREC 2004, Lisbon, Portugal, pp. 1553–1556 (2004)
2. Clavel-Merrin, G.: MACS (Multilingual access to subjects): a virtual authority file across languages. *Cataloging & Classification Quarterly* 39(1/2), 323–330 (2004)
3. Humphreys, B.L., Lindberg, D.A., Schoolman, H.M., Barnett, G.O.: The unified medical language system: an informatics research collaboration. *J. Am. Med. Inform. Assoc.* 5(1), 1–11 (1998)
4. Liang, A.C., Sini, M.: Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures. *New Review of Hypermedia and Multimedia* 12(1), 51–62 (2006)
5. Zeng, M.L., Chan, L.M.: Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology* 55(5), 377–395 (2004)
6. Chen, S.J.: Chinese-language Art & Architecture Thesaurus: methods and issues. In: CI-DOC 2009, Chile (September 2009), <http://aattaiwan.teldap.tw/2009/10/cidoc-2009.html>
7. Chen, S.J., Wu, D.: Contribution and creation of new concepts in the bilingual thesaurus: methods and practices. In: AAT in Chinese Working Meeting. Getty Research Institute, L.A., USA (August 2009), <http://aattaiwan.teldap.tw/2009/09/aat.html>
8. Chen, S., Wu, D., Chang, Y.T.: Bilingual equivalence mapping: methods and issues. In: AAT in Chinese Working Meeting. Getty Research Institute, L.A., USA (August 2009), <http://aattaiwan.teldap.tw/2009/09/aat.html>
9. Cheng, C., Chen, Y.N., Chen, S.J., Chung, F.C., Lin, Y.H., Wu, D.: AAT-Taiwan System: initiative and progress. In: AAT in Chinese Working Meeting, Getty Research Institute, L.A., USA (August 2009), <http://aattaiwan.teldap.tw/2009/09/aat.html>

⁵ Chinese WordNet, <http://cwn.ling.sinica.edu.tw/>

Using Pattern Language as a Framework for Future Metadata Structure

Esben Agerbæk Black

The State and University Library, Denmark
Victor Albecks Vej 1, 8000 Århus C
eab@statsbiblioteket.dk

Abstract. In the 1970's Christopher Alexander envisioned the "pattern language". It contains an underlying philosophy [1] of what to accomplish by using pattern language; it is this philosophy we tap into and apply to metadata planning.

Different collections needs different metadata to be of future use; this information has a structure, we aim to reuse knowledge of, and standerize the creation of these structures. We further believe pattern language will ease the transition of existing digital collections.

1 Introduction

Collections are chosen, analysed and digitized, in a primarily one-way process which offers little opportunity to make revisions the collection metadata. So we must get it right the first time, or at the very least be willing to repeat the process of selecting and applying metadata.

The process of selecting and structuring the metadata is a process combining the work of collection experts, librarians and technicians. Limitations in the availability of personnel, opportunities for collaboration and the resources available, make an approach towards optimizing the decision process important.

2 Two Kinds of Metadata, Two Different Approaches

When it comes to metadata we deal with two subsets, of which the first is technical and primarily machine generated. We include as much machine generated metadata as possible.

The other subset is the descriptive metadata. Each entry may have different metadata. This is where metadata planning becomes challenging. The descriptive metadata will vary greatly between collections – consider having to describe the combined metadata of LP-records as audio-books, speech recordings, audio plays and music.

More importantly there will need to be some degree of human control and/or work involved. This naturally increase the costs, choices must be made in regard to what and in which detail descriptive metadata is included.

3 The Philosophy of Pattern Language

“First, it has a moral component. Second, it has the aim of creating coherence, morphological coherence in the things which are made with it. And third, it is generative: it allows people to create coherence, morally sound objects, and encourages and enables this process because of its emphasis on the coherence of the created whole.” [1]

It is imperative that work instructions are coherent and have a concise form. Achieving this in many patterns over a wide spectrum of subjects is not a trivial task, and have resulted in an instruction for the use and creation of patterns. The instruction is inspired by the work of C. Alexander et al. [2] and shown in section 3.2. A conceptual relation between natural languages and the structure chosen here is shown in figure 1.

3.1 Structural Rules for Patterns

The users of the metadata pattern language should be able to create new patterns to add to already existing patterns. augmenting and adding patterns, expanding the language vocabulary.

The adding of a pattern must adhere to a common structure, as described in section 3.2, and add value, consistency and expressive power to the language as a whole. [3]

A pattern is meant to shed light on a best practice for a given problem domain or situation that is complex.

3.2 Patterns Described

A pattern has a structure, here represented by an exert from the library’s internal documentation.

Grade: from 0(zero) to 2(two) stars. -** appended to the Name, the grade is an indication of the pattern quality and completeness

Name (and grade): Try to use a descriptive name, make it catchy for ease of memory, but not so catchy that it becomes too creative.

Context: Where in the process does this pattern apply? Remember to include relevant patterns in-line in the context description.

Description of pattern: The general description, it provides no problems or solutions, just a description of the covered subject. *Section should be written in bold font*

Problem description: A description of the relevant forces and constraints, and how they interact.

Solution: What is the encouraged approach to solving the problem of this pattern? *Section should be written in bold font*

Consider next: What are the options provided by this pattern, and what direction does it encourage?

A single pattern, on its own, offers little meaning in the creation of collections; it is meant to be used to form expressions that can be used as a tool for articulating the structure of the implemented collections. A structural analogy to natural languages is shown in figure [II](#)

Natural language	Pattern language
Words	Patterns
Rules of grammar and meaning which express connections	Pattern expressions
Sentences	Implemented collections

Fig. 1. Relation between natural and pattern language

3.3 An Example from the Women’s History Archive

The introduction of a new content model [4](#) description will form the example in this case. We will look at the example of a collection of *grey material*, mostly flyers and members’ bulletins during the period from the late nineteenth century to present day. Only a few subcollections have been digitized so far.

This requires the use of a pattern describing the re-use of old collections. The pattern already exists, and is available as a description of the inclusion of a generic old collection.

The following is an example of such a pattern for *grey material*.

Grey material **

This pattern applies to *grey material* collections or sub collections of *grey collections*, and is relevant to other small-set print collections with limited available metadata. The material is often of irregular size and print-quality and the subject of the content is often very varied. It could be protest lyrics, diary entries or personal recounts to political manifestations.

The grey material pattern applies to small-set print productions, such as those made on duplicator machines and applies especially in relation to digitization processes for existing analogue *grey material* collections.

A record of only standard Dublin Core entries will not always be sufficient, since relations to the social group behind the material could be advantageous to include where possible, as could the possible use of keywords.

The great variance of the data, both in content and existing metadata is problematic when considering the structure of the grey material collection.

When using this pattern strive to include the *Dublin Core* pattern and attempt to extend this by using *DC qualifiers*. For textual grey material include the *article* pattern from the newspaper patterns.

Consider the inclusion of the *Dublin Core* pattern and the *Archive description* pattern, and as always remember to scrutinize your work, since there might be subtle relations to other textually oriented pattern, i.e. *Articles*, *Story telling* and others. Remember to adhere to the *Cost efficiency* principle described in *Metadata for Digital Resources* [5].

4 Integrating Different Metadata Approaches

One of the efforts of our approach to metadata creation has been to include other metadata systems in our patterns, for instance the *FRBR* [6] standard as well as the *Dublin Core* [7] elements. The FRBR Relations have been implemented as a pattern [8].

5 Conclusion

So far we have had good experience with the use of pattern language in relation to deciding on metadata structure. When working with existing collections the knowledge and patterns of previous conversion tasks becomes a guideline and a supportive structure for the work at hand.

In the generation of new metadata patterns the availability of existing patterns aid in the creation of new patterns and the fullness of the expressions, both as components of patterns as well as inspiration for new patterns. Boosting the creation of new metadata structures.

References

1. Alexander, C.: The origins of pattern theory: the future of the theory, and the generation of a living world. *IEEE Software* 16(5), 71–82 (2002)
2. Alexander, C., Ishikawa, S., Silverstein, M.: *A Pattern Language: Towns, Buildings, Construction*. Center for Environmental Structure Series, Later printing edn. Oxford University Press, Oxford (1977)
3. Alexander, C.: *The Timeless Way of Building*. Oxford University Press, Oxford
4. DOMS Content Model, http://wiki.statsbiblioteket.dk/domswiki/DOMS_Content_Model
5. Foulonneau, M., Riley, J.: *Metadata for Digital Resources: Implementation, Systems Design and Interoperability*. Chandos Publishing, Oxford (2008)
6. Functional Requirements for Bibliographic Records - IFLA Cataloguing Section, http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm
7. DCMI Home: Dublin Core® Metadata Initiative (DCMI), <http://dublincore.org/>
8. FRBR Relations – Domswiki, http://wiki.statsbiblioteket.dk/domswiki/GuidelinesForNewDatamodel/PatternLanguage/FRBR_Relations

i-TEL-u: A Query Suggestion Tool for Integrating Heterogeneous Contexts in a Digital Library

Maristella Agosti, Davide Cisco, Giorgio Maria Di Nunzio,
Ivano Masiero, and Massimo Melucci

Department of Information Engineering – University of Padua
Via Gradenigo, 6/a – 35131 Padua – Italy

Abstract. This paper presents the design, implementation and evaluation of a query suggestion tool (named i-TEL-u) that allows for the management and the exploitation of different contexts in an integrated way within the same search interface for accessing the contents of The European Library portal¹. i-TEL-u allows users to seamlessly move from one context to another according to their information needs and to the way these needs evolve during the search session. The aim of this tool is to improve the search functionalities of the portal, attract many users and give them easy and effective access².

1 Problem and Contribution

Years of observation of Web search engine usage have shown that people do not express their needs by means of verbose natural language sentences, rather they tend to summarize them with two or three words on average. This user behavior poses a major challenge to search engines whose effectiveness largely depends on how queries are posed and, in particular, on the number, clarity and specificity of the query words. In order to tackle this problem, search engines facilitate access by adding tools to their interfaces, such as query suggestion tools, to enhance the search experience, gather user behavior information, improve effectiveness, and provide users with relevant results. [1,2,3,4,5,6,7,8]

TEL is a free service that offers access to the resources of 48 national libraries of Europe in 35 languages, provides a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage. Resources can be both digital (e.g. books, posters, maps, sound recordings, videos) and bibliographical; the quality and reliability of the documents are guaranteed by the collaborating national libraries. The TEL users are “average” Internet users and expect from TEL the search effectiveness and usability offered by search engines. Basically, TEL offers two types of search, i.e. the simple

¹ <http://www.theeuropeanlibrary.org/>

² The work was partially supported by TELplus Targeted Project - eContent*plus* Program of the European Commission (Contract ECP-2006-DILI-510003).

search and the advanced search like many library portals. The simple search allows the users to search the TEL collection using keywords and users exploit these functionalities like they do with a search engine, i.e. they type two or three words and browse the top results. However, the search results returned by TEL in this way are rather unsatisfactory not only due to the queries, but also due to the type of documents stored, which are library records.

The main idea underlying our contribution is to provide users with search functionalities similar to those they deem effective when using search engines to retrieve pages from the Web, that is, a search box available on the Web browser and, most importantly, a query suggestion tool focussed on and tailored to TEL. This tool was named i-TEL-u and it leverages a variety of data sources for implementing different contexts while traditional query suggestion tools cannot recognize different contexts from which the suggested queries are produced.

i-TEL-u allows users to seamlessly move from one context to another according to their information needs and to the way these needs evolve during the search session. As the user types the query, a list of terms extracted from one of the contexts available is shown to the user who can easily move from one context to another with a slider in the search interface, thus getting new suggestions from the other contexts³.

2 Design of i-TEL-u

The approach taken to design i-TEL-u was based on a notion of context as “wisdom of the crowds” made available by the TEL users experiences and the Web search engines. Context has thus been conceived as the “dataspace” where the data are observed and, subsequently, it acts as the source from which suggestions are computed. A model which combines data coming from three main dataspace through which the knowledge moves from a user-centered to a broader context was defined. Specifically, the following main dataspace and algorithms were defined and implemented: **Lexical**, co-occurrence and frequency data taken at run-time from the most popular queries of two major search engines, completing terms as commercial search engines do - used in the Web Search Context; **Semantic**, term and relationship representation taken from a world-wide ontology built on top of Wikipedia, finding related topics rather than related terms - used in the Wiki Context; **Statistical**, the frequency distribution of pairs of terms stored in the TEL log files, using term distribution - used in the TEL Context. We also investigated the possibility of merging results of two dataspace: the Global Context which combines the Lexical and the Semantic dataspace was analyzed during the user study.

3 User Study

The user study consisted of two steps: the collection of a set of candidate queries, and the evaluation of the suggestions concerning the tool based on this set of

³ <http://ims.dei.unipd.it/websites/TelPlus/iTELUVideo.m4v>

Table 1. For each user and context we reported on the first line the number of queries suggested by i-TEL-u with at least one record, between brackets the number of records judged not relevant/partially relevant/relevant, on the second line the averaged NDCG

user	measure	Web	Global	Wiki	TEL
user 1	suggestions	4 (6/9/29)	1 (2/2/8)	4 (29/14/7)	4 (21/9/9)
	NDCG	0.6737	0.1638	0.3553	0.4357
user 2	suggestions	4 (24/1/7)	1 (7/0/2)	3 (16/5/9)	3 (8/7/6)
	NDCG	0.2553	0.0808	0.2417	0.2420
user 3	suggestions	2 (7/9/3)	0 (0/0/0)	1 (15/0/0)	5 (42/3/6)
	NDCG	0.2380	0.0000	0.0000	0.1517
user 4	suggestions	2 (1/4/15)	1 (0/2/1)	2 (10/8/2)	5 (14/5/17)
	NDCG	0.3491	0.1380	0.1264	0.5933
user 5	suggestions	4 (8/24/5)	2 (2/7/2)	3 (16/26/2)	4 (7/19/5)
	NDCG	0.4180	0.1285	0.2447	0.3763
user 6	suggestions	4 (15/10/7)	5 (34/9/7)	2 (13/4/3)	4 (22/3/0)
	NDCG	0.3660	0.1883	0.1173	0.0307
Total	suggestions	20 (61/57/66)	10 (45/20/20)	15 (99/57/33)	25 (114/46/43)
	NDCG	0.3834	0.1166	0.1809	0.3049

queries. During the first step, the seven users who participated in this study were asked to navigate the TEL portal, become familiar with the search interface and find five queries that did not have any result (no records found in TEL). In the next phase, we wanted to study whether the tool could suggest a new query able to retrieve a number of records different from zero, and whether these records were relevant for the user for each context.

During the second phase, the 35 queries collected were redistributed to the participants with the following strategy: of the five queries given to each user, three queries were originally prepared by himself, the other two queries were randomly collected from the other queries prepared by someone else. For each query users were asked to judge the results proposed by each context of i-TEL-u and rate the retrieved records with one of the three possible values: not relevant, partially relevant, relevant. We used the normalized discounted cumulative gain (NDCG) to measure the performance of the suggestions given by i-TEL-u. Six of the seven users participated in this phase; therefore, a total of 30 queries were evaluated.

In Table 1 for each user and for each context we reported the number of suggested queries that retrieved at least one record, the number of records judged ‘not relevant’/‘partially relevant’/‘relevant’, the NDCG, and in the last row the overall sum. The Web context and the TEL context emerge over the other two in terms of number of queries with at least one suggestion and NDCG (respectively 20 and 25, and 0.3834 and 0.3049). This is an important result since it shows that two orthogonal contexts - suggestions given by major search engines and correlations among search terms in the TEL logs - can help the user in refining and finding results not reachable originally. It is important to underline the fact that the TEL context gives more suggestions but of poorer quality as emerges

from the NDCG. The number of suggestions that are judged positive is 66 for the Web context, equal to 36% of the suggestions, 20 for the Global context, equal to 24% of the suggestions, 33 for the Wiki context, equal to 17% of the suggestions, and 43 for the TEL context, equal to 21% of the suggestions. This means that on average a user can expect to get a positive suggestion every 4 or 5 suggestions given by the tool; this distribution is approximate, since it is more likely to have very difficult queries with very few or no good suggestions and others which get very good scores for many documents. The number of times the tool gives suggestions which are all judged negative is small: once over 20 for the Web context, once over 15 for the Wiki context, and five times over 25 for the TEL context; this can be interpreted in the following way: if the tool gives the user at least one suggestion, it is very likely that the suggestion can be at least of partial help.

A sign test was performed to understand whether the difference in the gain in terms of NDCG were statistically significant compared to the situation of not having any help from the tool. We considered the distribution of the 30 values of NDCG separately for each context and performed the test to verify whether the distribution came from a distribution with zero median and alpha value equal to 5%. In all four contexts, the null hypothesis had to be rejected with p-values very close to zero (apart from the “Global” context which presents a p-value equal to 0.003 that could be worth a deeper analysis), which means that the distribution of the results is significantly different (in positive) from a distribution which has all zeros and median zero which is the case of not using the tool.

References

1. Efthimiadis, E.: Query expansion. In: Williams, M. (ed.) *Annual Review of Information Science and Technology (ARIST)*, vol. 31, pp. 121–185. Information Today for the American Society for Information Science, Medford (1996)
2. Lalmas, M., Ruthven, I.: A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review* 18(1) (2003)
3. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: *KDD*, pp. 875–883. ACM, New York (2008)
4. Ma, H., Yang, H., King, I., Lyu, M.R.: Learning latent semantic relations from clickthrough data for query suggestion. In: *CIKM*, pp. 709–718. ACM, New York (2008)
5. Mei, Q., Zhou, D., Church, K.: Query suggestion using hitting time. In: *CIKM*, pp. 469–478. ACM, New York (2008)
6. Gao, W., Niu, C., Nie, J.Y., Zhou, M., Hu, J., Wong, K.F., Hon, H.W.: Cross-lingual query suggestion using query logs of different languages. In: *SIGIR*, pp. 463–470. ACM, New York (2007)
7. Huang, C.K., Chien, L.F., Oyang, Y.J.: Relevant term suggestion in interactive web search based on contextual information in query session logs. *JASIST* 54(7), 638–649 (2003)
8. White, R.W., Marchionini, G.: Examining the effectiveness of real-time query expansion. *Inf. Process. Manage.* 43(3), 685–704 (2007)

The Planets Testbed

A Collaborative Research Environment for Digital Preservation

Brian Aitken¹, Seamus Ross¹, Andrew Lindley², Edith Michaeler³,
Andrew Jackson⁴, and Maurice van den Dobbelsteen⁵

¹ Humanities Advanced Technology and Information Institute George Service House
11 University Gardens, University of Glasgow, Scotland, G12 8QH, UK
`b.aitken@hatii.arts.gla.ac.uk`, `s.ross@hatii.arts.gla.ac.uk`

² Austrian Institute of Technology, Donau-City-Strasse 1, 1220 Vienna, Austria
`andrew.lindley@ait.ac.at`

³ Austrian National Library, Josefsplatz 1, A-1015 Vienna, Austria
`edith.michaelar@onb.ac.at`

⁴ The British Library, Boston Spa, West Yorkshire, LS23 7BQ, UK
`andrew.jackson@bl.uk`

⁵ National Archives of the Netherlands, The Hague, The Netherlands
`maurice.van.den.dobbelsteen@nationaalarchief.nl`

Abstract. The digital objects that are so fundamental to 21st century life may have a precarious future due to the rapid pace of technological change. Digital preservation specialists have proposed an almost overwhelming variety of preservation actions and tools that may help to mitigate this risk, but there is a lack of empirical evidence to help librarians, archivists and non-specialists to make an informed decision about the most applicable and effective preservation tools. The Planets project has developed a digital preservation Testbed that aims to provide such an evidence-base.

The Planets Testbed (<http://testbed.planets-project.eu/testbed/>) is a freely available and easy to use controlled environment where users can experience and compare different preservation tools and approaches through their web browser. The Planets approach is to make preservation tools available through a service-oriented architecture; the tools, which may run on any platform, are given a web service wrapper which then allows users to access certain aspects of the tool's functionality from within the Testbed's web-based interface, as shown in Figure 1. Through the Testbed users can design and execute a variety of experiments, such as migration and emulation experiments. The focus of a migration experiment may be to analyse the performance and trustworthiness of tools that transform digital objects from one format (such as obsolete word processor files) into more up-to-date or preservable formats (such as PDF/A). The focus of an emulation experiment may be to investigate how effectively and accurately an obsolete digital object is rendered within an emulated hardware and software environment.

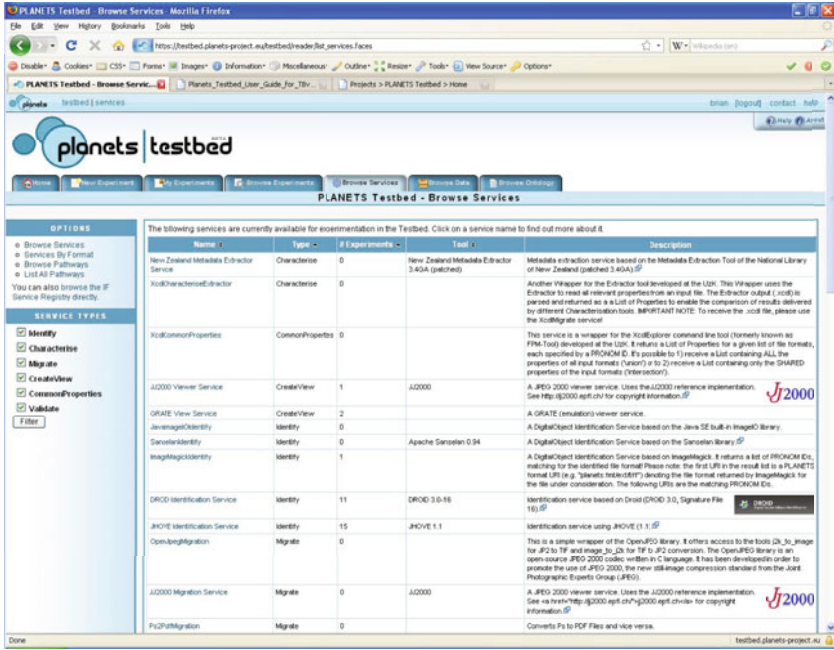


Fig. 1. The Testbed Interface, listing services

In order to perform such experiments digital objects must be passed to the preservation tool for processing, and users can experiment on their own data using a dedicated FTP upload area. Additionally, the Testbed provides access to several large corpora of test files that cover a variety of popular and important file formats. These include edge-case files such as malformed PDFs, GIF files that have experienced bit-rot and files that have been given the wrong extension. The Testbed not only provides access to this vast array of sample data, but the properties of the corpora data have been documented using the Planets-developed Extensible Characterisation Definition Language (XCDL), which makes them ideal control files for experimentation.

One of the principal aims of the Testbed is to create a shared knowledge-base of digital preservation tool performance, and for this reason experiment details, input files and outcomes are made available to all Testbed users. If a user is interested in migrating a collection of GIFs or would like to know how effective OpenOffice is at generating PDFs from Word 97 files then the database of existing experiments can be browsed for such experiments. Furthermore, the Testbed facilitates the reproducibility of experiments: users can re-run any experiment to prove the validity of the results, or even adapt an existing experiment to fulfil new requirements. In addition users can rate and review their own and other users' experiments, supplying comments and engaging in discussions through the Testbed interface.

Testbed experiments follow a simple to use yet flexible six-step process, as shown in Figure 2. At Step 1 the basic properties for the experiment are defined, including the overall experiment aims and objectives, contact details and references. In Step 2 the user formulates the design of the experiment, which includes selecting an experiment type, choosing the required preservation services and tool parameters, and attaching the data that will be experimented upon. Step 3 is where the experiment is executed. At this point the input files are processed by the selected services, statistics relating to the execution are gathered and (if applicable) output files are generated and returned to the Testbed. At

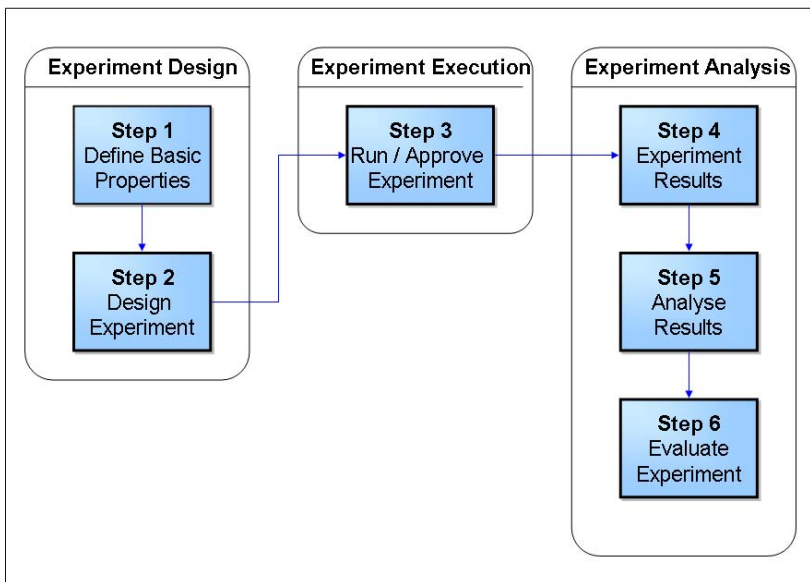


Fig. 2. The 6-step Experiment Process

Step 4 the results of the experiment are displayed. Input and output files are listed, and can be opened or downloaded if required. A selection of file properties such as filename, file size and (in the case of images) thumbnails are displayed and execution time per digital object is displayed graphically. Additionally, this page displays whether the operation conducted on each input file was a success.

In Step 5 the results can be analysed. In order to assess the effectiveness of a preservation tool we need to investigate how digital objects that have undergone a preservation action differ from their original form. A variety of characterisation and identification services that can automatically extract digital object properties are offered at this stage, together with options for manually measuring properties. By comparing the significant characteristics in the original objects with the post-preservation action objects (i.e. migrated files or objects within

an emulated environment) it is possible to gain an understanding of the effectiveness of a tool. Finally, in Step 6 the user can provide an overall evaluation of their experiment, stating how effectively the tools preserved the properties and any other factors the user wishes to document.

By supplying a controlled environment where users can experiment with preservation tools and share outcomes with others, the Planets Testbed aims to enhance the knowledge of preservation approaches and help users to make informed decisions about which tools and approaches are the most useful for particular tasks.

Acknowledgements

The Planets Testbed Research and Development work presented within this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD - Project IST-033789.

A Functionality Perspective on Digital Library Interoperability

George Athanasopoulos¹, Edward Fox², Yannis Ioannidis¹, George Kakalettris¹,
Natalia Manola¹, Carlo Meghini³, Andreas Rauber⁴, and Dagobert Soergel⁵

¹ Dept. of Informatics and Telecommunications, University of Athens, Greece
{gathanas,yannis,gkakas,natalia}@di.uoa.gr

² Dept. of Computer Science, Virginia Tech, Blacksburg, VA, USA
fox@vt.edu

³ Istituto della Scienza e delle Tecnologie della Informazione,
Consiglio Nazionale delle Ricerche, Pisa, Italy
carlo.meghini@isti.cnr.it

⁴ Dept. of Software Technology and Interactive Systems,
Vienna University of Technology, Vienna, Austria
rauber@ifs.tuwien.ac.at

⁵ Dept. of Library and Information Studies, University of Buffalo, New York, USA
dsoergel@buffalo.edu

Abstract. Digital Library (DL) interoperability requires addressing a variety of issues associated with functionality. We report on the analysis and solutions identified by the Functionality Working Group of the DL.org project during its deliberations on DL interoperability. Ultimately, we hope that work based on our perspective will lead to improved architectures and software, as well as to greater interoperability, for next-generation DL systems.

1 Introduction

A huge volume of information and knowledge is acquired and managed by distinct Digital Libraries (DLs). This leads to problems for academic and public libraries that often work with scores of such DLs and seek to support patrons facing a broad range of systems and services. Similar problems are faced by students, faculty, researchers, scholars, knowledge workers, and the general public. Also of concern is e-science, where labs and centers must use different DLs to address global challenges.

Interoperability among all the DLs needed in each case is a serious concern. Manifesting a broad range of features and capabilities, DL systems employ diverse proprietary solutions and varying applications of a broad range of standards. The problem is further aggravated by the complexity and scale of modern DL systems and problems such as API mismatch, data format mismatch, and missing components.

Interoperability has been the main issue of concern for the DL.org project [4]. Its work is based on the DELOS Digital Library Reference Model [3], in particular, the multi-dimensional representation of the DL domain and the identification of six primary concepts that characterize Digital Libraries: content, users, functionality, policy, quality, and architecture. In this paper, we present results from the discussions

of the DL.org Functionality Working Group [4]. This Working Group is focusing on interoperability with respect to one of these concepts, DL functionality, while still remaining within the broader context of the entire DL space.

2 Functions, Interoperability, and Compatibility

There are many definitions of *function*. The Reference Model defines function as: “*an action a DL component or a DL user performs*” [3]. In software engineering, a function is a portion of code or a module that performs a specific task (or action); it is embedded within a larger program but remains relatively independent from the rest of the code. Function interoperability is often concerned with software modules that implement a DL function. The mathematical definition of a function as a mapping from a domain to a range is not important here.

Function interoperability is of particular importance in DLs, as indicated in Table 1. Such interoperability serves three main purposes: i) To provide users of one DL access to the content and functionality of other DLs; ii) To harmonize the user experience provided by different DLs so that the user who has learned to use one DL can use other DLs easily; iii) To save effort in creating new DLs or adding functionality to existing DLs, by reusing existing software components.

Table 1. Indicative set of functions where interoperability is especially important

Behind the scenes	For users
Feature Extraction	Federated search.
Classification/Clustering	Integration of additional external content sources on the fly.
Single Authorization/Single Sign-on	Visualization of timelines, maps, videos, etc.
Analysis/Statistics operations	Browsing based on same look-and-feel.
User Profile Management	
Harvesting, Aggregating	
Preservation and Backup	

To achieve interoperability of functions, one may use a registry that allows for the discovery of software modules that implement sought-after functionality in a given software context for a given user group. Such a registry should show the different ways in which functions can be interoperable.

From a system-based point of view, three important ways to achieve function interoperability are the following:

1. Based on processing (e.g., function Fa may utilize the functionality offered by another function Fb either by directly incorporating the provided functionality within Fa or by calling it as an external service);
2. Based on data/content (e.g., the outcome of function Fa is da which is used by Fb as input via direct exchange / conversion / replacement);
3. Based on cross-function compatibility (e.g., functions Fa and Fb have the same interface).

From a user point of view there are issues related to functionality that concern the compatibility of products. Such compatibility is directional and can be expressed in

two levels. More formally, DL B is product-compatible with DL A with respect to functionality if the following hold:

1. DL B provides all functions that DL A provides (product compatibility with respect to functions provided),
2. DL B uses the same interface with DL A to invoke functions, where interface refers to names of functions, shape and color of buttons, screen layout, command names, and syntax (full product compatibility).

From a framework based point of view, function interoperability requires:

1. an entity-relationship schema,
2. a taxonomy of the ways in which functions can interoperate, and
3. a template for the description of functions and software components.

3 Specifications, Solutions, and Ontologies

A function specification, using a template such as the following, facilitates identification of what a function does and how a system or a human may interact with it.

- **Function Behavior:** providing a description of what a function does and of the supported interaction with actors (systems/users)
- **API/Interface Specification:** illustrating the Input and Output data and parameters, data formats/standards, pre-conditions and post-conditions
- **Dependencies/Relationships/Use:** detailing the operating environment in which a function runs; other functions needed; functions that invoke this function; composite functions and workflows
- **Interoperability Concerns:** documenting what is required for interoperability and how does a specific implementation meet these requirements
- **Performance evaluation, assessment, and advice for use**

Specifications of many of the above properties can use existing widely used standards such as IDL, WSDL, SAWSDL, OWL-S, WSMO, or BPEL4WS.

Clearly, there are generic concerns regarding interoperability of functions that cut across all types of software systems and others that are particular to DLs. A function can be implemented as a *service*; thus the Service Oriented Computing (SOC) domain is of particular interest. In that context, there are already proposed static and dynamic solutions. Static solutions can be found in the e-Framework (www.e-framework.org) or the RosettaNet community (www.rosettanet.org). Both initiatives accommodate a standards-based service-oriented approach with well-defined services that facilitate the whole range of the functionality and provided features. On the other hand, dynamic solutions address several of the specified issues and rely on the use of formally defined theories for the automated provision of adapters. Adapters handling incompatibilities (e.g., with interfaces, behavior, and pre/post conditions) have been developed by the SOC research community [1, 2]. These two distinctive types of solutions can serve as the basis for DL function interoperability as well.

Another important aspect when dealing with interoperability is that functions can be related in various ways, including the common relationship of sub-function. As Table 2 shows, the standard DL function “search” has many sub-functions, especially

when advanced search is concerned. Thus, it is essential that taxonomic or, even better, ontological descriptions of DL functions be provided. A thorough treatment of this matter is given elsewhere [5], using the 5S framework; other work involves ontology mappings, alignments, and merging [6].

Table 2. Sub-functions of search

Quick Search	Advanced Search
Enter a query and click search	Enter keywords or phrases for selected fields
Enter keywords or phrases for selected field	Select keyword from a list
Limit results	Select Boolean operator (explicit)
Search subscribed titles	Define phrase match (explicit)
Clear	Search within results
	Limit results to (preselection), Sort by (preselection)
	Select display options, Display X results per page
	Display search history

4 Conclusions

The Functionality Working Group of the DL.org project has explored issues, approaches, and solutions related to the interoperation of DL functions. Essential are appropriate description mechanisms and registries that will facilitate the publication and discovery of functions. More work is also needed on function taxonomies and ontologies, function composition, and the relationship to interoperability of the other DL concepts, as addressed by other DL.org Working Groups.

Acknowledgments. Partially supported by the European Commission under project “DL.org: Digital Library Interoperability, Best Practices and Modelling Foundations”, Contract num: 231551.

References

1. Benatallah, B., Casati, F., Grigori, D., Nezhad, R., Toumani, F.: Developing Adapters for Web Services Integration. In: Pastor, Ó., Falcão e Cunha, J. (eds.) CAiSE 2005. LNCS, vol. 3520, pp. 415–429. Springer, Heidelberg (2005)
2. Bordeaux, L., Salaün, G., Berardi, D., Mecella, M.: When are two Web Services Compatible? In: 5th VLDB Workshop on Technologies for E-Services, Toronto, Canada (August 2004)
3. Candela, L., Castelli, D., et al.: The DELOS Digital Library Reference Model. In: Foundations for Digital Libraries, Ver 0.98., DELOS Network of Excellence - Project no. 507618 (2008)
4. DL.org: Digital Library Interoperability, Best Practices and Modelling Foundations. EU funded project, Contract no. 231551, <http://www.dlorg.eu>
5. Goncalves, M.A., Fox, E.A., Watson, L.T.: Towards a Digital Library Theory: A Formal Digital Library Ontology. *Int. J. Digital Libraries* 8(2), 91–114 (2008)
6. Noy, N.: Semantic Integration: A Survey of Ontology-based Approaches. *ACM SIGMOD Record* 33(4), 65–70 (2004)

Overview and Results of the INEX 2009 Interactive Track

Thomas Beckers¹, Norbert Fuhr¹, Nils Pharo²,
Ragnar Nordlie², and Khairun Nisa Fachry³

¹ University of Duisburg-Essen, Germany
{tbeckers,fuhr}@is.inf.uni-due.de

² Oslo University College, Norway
{nils.pharo,ragnar.nordlie}@jbi.hio.no

³ University of Amsterdam, The Netherlands
k.n.fachry@uva.nl

Abstract. We present results of the INEX 2009 Interactive Track which focussed on how users behave in interactive search systems. Three types of working tasks based on a collection of book metadata were regarded. The results show differences with respect to the task types and point out improvements and new research questions for the next track in 2010.

1 Introduction and Research Questions

The INEX Interactive Track (iTrack) is a cooperative research effort run as part of the INEX Initiative for the Evaluation of XML retrieval [1]. The overall goal of the iTrack is to investigate how users behave in interactive search systems. In the 2009 run of this track [1] the focus was on what aspects of documents the users are interested in, how they make use of various search tools, and finding out new challenges for the coming iTracks. We created a collection based on a crawl of 2.7 million records from the book database of the online bookseller Amazon.com, consolidated with corresponding bibliographic records from the cooperative book cataloging web site LibraryThing.

In this paper, we provide an analysis of the collected logging and questionnaire data and point out challenges for the next run in 2010.

2 System Description

The search system (see fig. 1) was developed at the University of Duisburg-Essen. It is based on Daffodil [2] and partially on ezDL [2] while the retrieval component was implemented using Apache Solr. The client application part is based on a tool metaphor. Six tools (rounded rectangles in fig. 1) are available and are described in more detail below.

¹ <http://www.inex.otago.ac.nz/>

² <http://www.ezdl.de>, <http://www.is.inf.uni-due.de/projects/ezdl/>

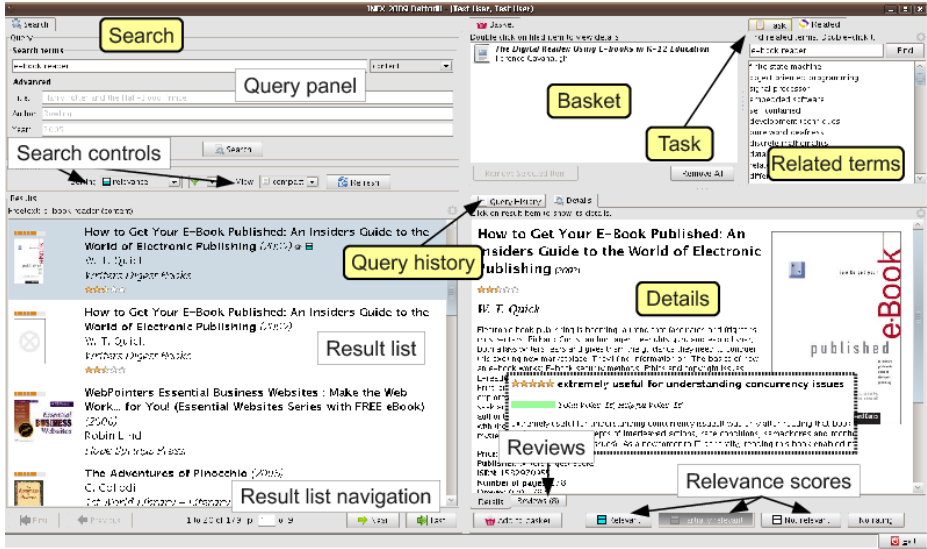


Fig. 1. The client desktop application of the iTrack search system

The **search tool** offers a Google-like search field as well as advanced search fields for title, author, and year. Below the query panel, the user can select fields for sorting or changing the display style of the result list. The lower half of the search tool contains the result list. The default surrogate contains the title, authors, year, publisher, average customer rating and a thumbnail of the book front cover. Each page of the result list contains 20 result items. The user can use the buttons at the bottom to navigate to other pages of the result list.

A double-click on a result item shows book details in the **detail tool**. Users can indicate the relevance of any book which is examined either as *Relevant*, *Partially relevant*, or *Not relevant*, by clicking markers at the bottom of the tool. A second tab shows reviews of the selected book.

The user can select any book as part of the answer to the search task by moving it to the **basket tool**. This can be done by drag-and-drop or by clicking the *Add to basket* button next to the relevance buttons.

A history of performed search queries is stored in the **query history tool**. The **related terms tool** presents terms related to those used in the search query. A search for related terms can also be triggered manually by the user. Finally, the **task tool** shows the current working task.

3 Evaluation Design

For this track, 41 volunteers were recruited mostly from students of computer science, cognitive and communication science, library science and some other related fields. 24 of them were male and 17 of them female. Their average age was

about 28. On average, they have used the Internet for 9.5 years. All participants had experiences with web search engines, searching in digital libraries or digital bookstores.

There are three different task groups, namely *broad tasks* (I), that require exploratory search behaviour, *narrow tasks* (II), that are about a relatively narrow topic, and a *self-selected* (III) task about finding a single book for a course the volunteers were currently attending. The first two task groups consist of three concrete working tasks of which one had to be chosen by the participant. Pharo et al. [1] provide a more detailed description of the working tasks.

At the beginning, the participants were asked to fill out a pre-experiment questionnaire. Each task was preceded by a pre-task and concluded by a post-task questionnaire. After all three working tasks had been worked on, a post-experiment questionnaire was answered. Actions by the system and the participants were recorded by the system and stored in a database.

4 Results

The participants were given the possibility to express positive as well as negative general comments on the questionnaires. The user interface was praised because it is well arranged and everything fits on a single screen without the need to scroll up or down. The inclusion of another document aspect, namely the reviews, was also pointed out positively by the participants.

Technical problems of the search system and sometimes useless related terms suggestions due to topical limitations of the data source were points of criticism. Participants also missed highlighting of query terms and filtering options for the result list. Also, the heterogeneity of the data was disliked, that is, some books have lots of metadata while other have just very little of it.

Table 1 shows the average event count per session for each task group. In task group I the details of books were requested more often than for other task groups while only slightly more searches were performed. The participants looked at the reviews for only a small part of the books they viewed details of. The related terms tool and the query history tool were used rarely. Less often than once per session a related term was searched/added to the query or a previous search was executed again.

Table 1. The average count per task session of the most important events

Event	I	II	III
Search	8.07	7.46	7.13
Next Search Results	2.66	1.93	2.63
View Details	23.78	15.44	16.05
View Review Details	3.66	4.61	3.35
Relevance Rating	15.66	9.39	9.33
Added To Basket	7.02	5.34	2.38
Sorting and View Changed	1.68	0.56	1.15
Related Terms Search	0.44	0.27	0.08
Related Term Selected	0.85	0.17	0.25
Query from History Selected	0.41	0.61	0.65

The average length of a session decreased from 829s (I) and 725s (II) to 622s (III). The average duration of a search (that is, the time between two queries) was 99s (I), 92s (II), and 83s (III). For more explorative working tasks, longer and more searches were performed.

At the end of a session the participants had 6.61 (I), 4.96 (II), and 1.15 (III) books respectively in their basket. The participants collected more books for the broad tasks than for the narrow tasks. The average query length (number of terms in the simple query field) was 1.99 (I), 2.46 (II), and 2.21 (III). The broadest tasks resulted in the shortest average query length.

The most frequent event transitions are similar for each task groups. The most frequent transition is *Visible Results Changed* → *View Details* which means that after results have been displayed or after the participant has scrolled the result list he/she views details of a book in the results. The second most frequent transition *View Details* → *Relevance Rating* was generated if the participants requested details and then assessed the relevance of the book with regard to the working task.

Overall, advanced tools such as the *related terms tool* or the *query history tool* were used less often than expected. It was often not necessary for the participants to use these tools to work successfully on the tasks. In our opinion the main reason is that the working tasks focussed too much on problems obvious from the title of a book. This was especially the case for task group III. Many participants understood this task group as a known-item search instead of a thoughtful search for a single unknown book as it was actually intended. The reviews were judged as a useful aspect but the participants mostly concentrated on the details only.

5 Conclusion and Outlook

The search system was well received by the participants. The use of metadata of real and recent books (until March 2009) from Amazon/LibraryThing was pointed out positively. Differences in the user behaviour in respect to different types of working tasks could be observed. This run of the iTrack has also shown some advices and challenges for the next run.

For the upcoming iTrack in 2010 we plan to design the working task such that they are expected to require more specific queries that cannot be matched by a whole book. The users will then be encouraged to not only rely on well-known metadata, such as title and author but also more on e. g. reviews or book covers. The book data is to be cleaned-up to increase the homogeneity, that is, entries with sparse metadata will be removed from the collection. Additionally, the search mode (*search* vs. *browse*) of users for the working tasks will be another research question. To investigate our hypotheses, we plan to perform a comparison of multiple systems.

References

1. Pharo, N., Nordlie, R., Fuhr, N., Beckers, T., Fachry, K.N.: Overview of the inex 2009 interactive track. In: Proceedings of the 8th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2009) (2010)
2. Fuhr, N., Klas, C.P., Schaefer, A., Mutschke, P.: Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, pp. 597–612. Springer, Heidelberg (2002)

SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size)

Jöran Beel^{1,2}, Bela Gipp^{1,2}, Ammar Shaker¹, and Nick Friedrich¹

¹ Otto-von-Guericke University, Computer Science/ITI/VLBA-Lab, Magdeburg, Germany

² UC Berkeley, Berkeley, California, USA

{beel, gipp, shaker, friedrich}@sciplore.org

Abstract. Extracting titles from a PDF's full text is an important task in information retrieval to identify PDFs. Existing approaches apply complicated and expensive (in terms of calculating power) machine learning algorithms such as Support Vector Machines and Conditional Random Fields. In this paper we present a simple rule based heuristic, which considers style information (font size) to identify a PDF's title. In a first experiment we show that this heuristic delivers better results (77.9% accuracy) than a support vector machine by CiteSeer (69.4% accuracy) in an 'academic search engine' scenario and better run times (8:19 minutes vs. 57:26 minutes).

Keywords: header extraction, title extraction, style information, document analysis.

1 Introduction

Extracting the title from PDF documents is one of the prerequisites for many tasks in information retrieval. Among others, (academic) search engines need to identify PDF files found on the Web. One possibility to identify a PDF file is extracting the title directly from the PDF's metadata. However, often the PDF metadata is incorrect or missing. Therefore, what is often tried is to extract the title from the PDFs' full text.

Usually, machine learning approaches such as Support Vector Machines (SVM), Hidden Markov Models and Conditional Random Fields are used for extracting titles from a document's full text. According to studies, the existing approaches achieve excellent accuracy, significantly above 90%, sometimes close to 100% [1, 2, 3]. However, all existing approaches for extracting titles from PDF files have two shortcomings. First, they are expensive in terms of runtime. Second, they usually convert PDF files to plain text and lose all style information such as font size.

For our academic search engine SciPlore.org we developed *SciPlore Xtract*, a tool applying rule based heuristics to extract titles from PDF files. In this paper we present this tool, the applied heuristics and an evaluation.

2 SciPlore Xtract

SciPlore Xtract is an open source Java program that is based on pdftohtml¹ and runs on Windows, Linux and MacOS. The basic idea is to identify a title based on the rule that it will be the largest font on the upper first third on the first page.

Google Scholar's Ranking Algorithm: An Introductory Overview

Jöran Beel
 Otto-von-Guericke University
 Department of Computer Science

Bela Gipp
 Otto-von-Guericke University
 Department of Computer Science

Fig. 1. Example PDF

```

5 <page number="1" position="absolute" top="0" left="0" height="1262" width="893">
6 <fontspec id="0" size="22" family="Times" color="#000000"/>
7 <fontspec id="1" size="16" family="Times" color="#000000"/>
8 <fontspec id="2" size="9" family="Times" color="#000000"/>
9 <fontspec id="3" size="7" family="Times" color="#000000"/>
10 <fontspec id="4" size="13" family="Times" color="#000000"/>
11 <fontspec id="5" size="13" family="Times" color="#000000"/>
12 <fontspec id="6" size="16" family="Times" color="#000000"/>
13 <text top="106" left="245" width="415" height="27" font="0">Google Scholars Ranking Algorithm: An </text>
14 <text top="134" left="336" width="231" height="23" font="0">Introductory Overview </text>
15 <text top="189" left="348" width="77" height="20" font="1">Jöran Beel</text>
16 <text top="186" left="424" width="6" height="13" font="2">1</text>
17 <text top="189" left="430" width="108" height="20" font="1"> and Bela Gipp</text>
18 <text top="186" left="538" width="6" height="13" font="2">2</text>
19 <text top="189" left="545" width="6" height="20" font="1"> </text>
20 <text top="225" left="340" width="7" height="11" font="3"><i>1 </i></text>
21 <text top="227" left="347" width="106" height="17" font="4"><i>JoeranBeel.org </i></text>
22 <text top="225" left="453" width="9" height="11" font="3"><i>2</i></text>
23 <text top="227" left="488" width="88" height="17" font="4"><i>BelaGipp.com </i></text>
24 <text top="244" left="201" width="496" height="17" font="5">Otto-von-Guericke University, Dept. of Comput
Science, Magdeburg, Germany </text>
25 <text top="289" left="106" width="71" height="20" font="6"><b>Abstract </b></text>
26 <text top="310" left="106" width="684" height="17" font="5">Google Scholar is one of the major academic
search engines but its ranking algorithm for academic articles is </text>
    
```

Fig. 2. Example XML Output

In the first step, SciPlore Xtract converts the entire PDF to an XML file. In contrast to many other converters, SciPlore Xtract keeps all layout information regarding text size and text position. Figure 2 shows an example XML output file of the PDF showed in Figure 1. Lines 6 to 12 of the XML file show all font sizes that are used in the entire document (in this case it is all “Times” in a size between 7 and 22 points). Below this, each line of the original PDF file is stated including layout information such as the exact position in which the line starts, and which font is used.

SciPlore Xtract now simply needs to identify the largest font type (in the example the font with the ID=0). Which text uses this font type on the first page is then identified and assumed to be the title.

3 Methodology

In an experiment, titles of 1000 PDF files were extracted with SciPlore Xtract. Then, titles from the same PDFs were extracted with a Support Vector Machine from CiteSeer [1] to compare results. CiteSeer’s tool is written in Perl and based on SVM Light² which is written in C. As CiteSeer’s SVM needs plain text, the PDFs were converted

¹ <http://www.pdftohtml.sourceforge.net>
² <http://svmlight.joachims.org/>

once with PDFBox³ and once with pdftotext⁴ as these are the tools recommended by CiteSeer. It was then checked for each PDF if the title was correctly extracted by SciPlore Xtract and CiteSeer's SVM (for both the pdftohtml text file and the PDFBox text file). If the title contained slight errors the title was still considered as being identified correctly. 'Slight errors' include wrongly encoded special characters or, for instance, the inclusion of single characters such as '*' at the end of the title.

The PDFs analyzed were a random sample from our SciPlore.org database, a scientific (web based) search engine. A title was seen as being correctly extracted when either the main title or both the main title and the sub-title (if existent) were correctly extracted. The analyzed PDFs were not always scientific. It occurred that PDFs represented other kind of documents such as websites or PowerPoint presentations. However, we consider the collection to be realistic for an academic search engine scenario.

4 Results

From 1000 PDFs, 307 could not be processed by SciPlore Xtract. Apparently, SciPlore Xtract (respectively pdftohtml) struggles with PDFs that consist of scanned images on which OCR has been applied. For further analysis only the remaining 693 PDFs were used. We consider this legitimate as the purpose of our experiment was not to evaluate SciPlore Xtract, but the applied rule based heuristic.

For 54 of the 693 PDFs (7.8%), titles could neither be extracted correctly by SciPlore Xtract nor CiteSeer's SVM. Only 160 (23.1%) of the titles were correctly identified by all three approaches. Overall, SciPlore Xtract extracted titles of 540 PDFs correctly (77.9%). CiteSeer's SVM applied to pdftotext identified 481 titles correctly (69.4%). CiteSeer's SVM applied to PDFBox extracted 448 titles correctly (64.6%). Table 1 shows all these results in an overview.

Table 1. Title Extraction of 693 PDFs

	Correct		Slight Errors		Total	
SciPlore Xtract	528	76.2%	12	1.7%	540	77.9%
CiteSeer SVM + pdftotext	406	58.6%	75	10.8%	481	69.4%
CiteSeer SVM + PDFBox	370	53.4%	78	11.3%	448	64.6%

When only completely correct titles are compared, SciPlore Xtract performs even better. It extracted 528 (76.2%) titles completely correct, while CiteSeer's SVM extracted only 406 (58.6%) respectively 370 (53.4%) completely correct.

SciPlore Xtract required 8:19 minutes for extracting the titles. SVM needed 57:26 minutes for extracting the titles from the plain text files (this does not include the time to convert the PDFs to text), which is 6.9 times longer. However, we need to

³ <http://pdftobox.apache.org/>

⁴ <http://www.foolabs.com/xpdf/download.html>

emphasize that these numbers are only comparable to a limited extent. CiteSeer's SVM extracts not only the title but also other header data such as the authors and CiteSeer's SVM is written in C and Perl while SciPlore Xtract is written in Java.

5 Discussion and Summary

All three tests show significantly worse results than the often claimed close-to-100% accuracies. Our tests showed (1) that style information such as font size is suitable in many cases to extract titles from PDF files (in our experiment in 77.9%). Surprisingly, our simple rule based heuristic performed better than a support vector machine. However, it could be that with other text to PDF converters, better results may be obtained by the SVM. CiteSeer states to use a commercial tool to convert PDFs to text and recommends PDFBox and pdftotext only as secondary choice. Our tests also showed (2) that runtime of the rule based heuristic was better (8:19 min) than SVM (57:26). However, these numbers are only limitedly comparable due to various reasons.

In next steps, we will analyze why many PDFs could not be converted (30.7%) and in which cases the heuristics could not identify titles correctly. The rule based heuristic also needs to be compared to other approaches such as Conditional Random Fields and Hidden Markov Models. We also intend to take a closer look at the other studies and investigate why they achieve accuracies of around 90%, while in our test the SVM achieved significantly lower accuracies. In the long run, machine learning algorithms probably should be combined with our rule based heuristic. We assume that this will deliver the best results. It also needs to be investigated how different approaches with different languages. Existing machine learning approaches mostly are trained with English documents. It might be that our approach will outperform machine learning approaches even more significantly with non-English documents as style information is language-independent (at least for western languages).

Summarized, despite the issue that many PDFs could not be converted, the rule based heuristic we introduced in this paper, delivers good results in extracting titles from scientific PDFs (77.9% accuracy). Surprisingly, this simple rule based heuristic performs better than a Support Vector Machine based approach.

Our dataset (PDFs, software, results) is available upon request so that other researchers can evaluate our heuristics and do further research.

References

- [1] Han, H., Giles, C.L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A.: Automatic document metadata extraction using support vector machines. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries, pp. 37–48 (2003)
- [2] Hu, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D., Zheng, Q.: Automatic extraction of titles from general documents using machine learning. *Information Processing and Management* 42(5), 1276–1293 (2006)
- [3] Peng, F., McCallum, A.: Accurate Information extraction from research papers using conditional random fields. *Information Processing and Management* 42(4), 963–979 (2006)

Academic Publication Management with PUMA – Collect, Organize and Share Publications

Dominik Benz¹, Andreas Hotho³, Robert Jäschke¹, Gerd Stumme¹,
Axel Halle², Angela Gerlach Sanches Lima², Helge Steenweg², and Sven Stefani²

¹ Knowledge & Data Engineering Group
University of Kassel & L3S Research Center
D-34121 Kassel, Germany

{lastname}@cs.uni-kassel.de

² University Library of Kassel

D-34127 Kassel, Germany

{lastname}@bibliothek.uni-kassel.de

³ University of Würzburg

Data Mining and Information Retrieval Group

Am Hubland, D-97074 Würzburg

{lastname}@informatik.uni-wuerzburg.de

Abstract. The PUMA project fosters the Open Access movement und aims at a better support of the researcher's publication work. PUMA stands for an integrated solution, where the upload of a publication results automatically in an update of both the personal and institutional homepage, the creation of an entry in a social bookmarking systems like BibSonomy, an entry in the academic reporting system of the university, and its publication in the institutional repository. In this poster, we present the main features of our solution.

Keywords: Publication Management, Puma, BibSonomy, Open Access, Institutional Repository, Tagging, Bookmarking, Metadata Sharing.

1 Introduction

The project „PUMA – Academic Publication Management“¹ is funded by the German Research Foundation (DFG) and has been started on August 1st, 2009. PUMA is a joint project of the University Library² and the Knowledge & Data Engineering Group³ of the University of Kassel (cf. [2]).

Open Access⁴ is a publication model that allows authors to publish their articles free of charge, and users to freely access them. The costs are borne by the institution that is providing the institutional repository. There are several reasons for this

¹ <http://puma.uni-kassel.de/>

² <http://www.ub.uni-kassel.de/>

³ <http://www.kde.cs.uni-kassel.de/>

⁴ <http://www.open-access.net/>

publication model. With reduced budgets and increased costs for journals, many university libraries cannot afford the subscription of all relevant journals any longer. Furthermore, Open Access supports a timely publication and broader visibility of articles so that research results can be taken up earlier and by more researchers, decreasing thus the turn around time of scientific results.

Even though many researchers support the open access movement in principle, they often do not contribute their publications to the institutional repository of their university. Key reasons are that they do not see an immediate benefit from this additional effort, and that the upload is not integrated in their usual workflow. PUMA aims therefore for an integrated solution, where the upload of a publication results automatically in an update of both the personal and institutional homepage, the creation of an entry in the social publication sharing platform BibSonomy,⁵ an entry in the academic reporting system of the university, and its publication in the institutional repository. At the time of upload, metadata from several data sources will be collected automatically in order to support the user. In addition, PUMA aims to provide a publication management platform for all researchers and students to be used on a daily basis, which reduces not only the open access publication effort but also the effort to manage one's own publications.

The PUMA is hosted by the University Library. It implements state-of-the-art Web 2.0 functionality. The platform includes all features known from BibSonomy, like tagging of publications, easy usage, an API and scalability. As a showcase, PUMA will be integrated with the open access repository platform DSpace, the library system PICA, the Typo3 content management system, and BibSonomy. The system is open for adaptation to other standards and systems. The project results will be published as open source software. This implies that any university resp. university library will be able to build its own publication management according to individual needs.

2 Architecture

PUMA is based on the well-known social bookmarking system BibSonomy, which allows users to organise and share bookmarks and publications in a collaborative manner. The basic building blocks of PUMA are an Apache Tomcat servlet container using Java servlet technology and a MySQL database as backend. All search engine like requests are handled by an adopted Lucene framework. A detailed description can be found in [1].

3 User Interface

The user interface is depicted in Fig.1, which shows bookmarks, publication posts and tags of a user. The page is divided into four parts: the header (showing information such as the current page title and location, navigation links and search boxes), two

⁵ <http://www.bibsonomy.org/>

lists of posts – one for bookmarks and one for publications – each sorted by date in descending order, and a list of tags related to the posts.



Fig. 1. PUMA displays bookmarks, publication metadata and tags simultaneously

4 Features

All possibilities known from BibSonomy are included within PUMA. In addition, there are some more advanced features: Apart from standard folksonomy features such as an intuitive user interface, navigation along all dimensions, or browser integration via RSS feeds, PUMA provides tag hierarchies, group management and privacy features, and numerous import and export functions, in particular to and from BibTeX, EndNote, and Zotero.

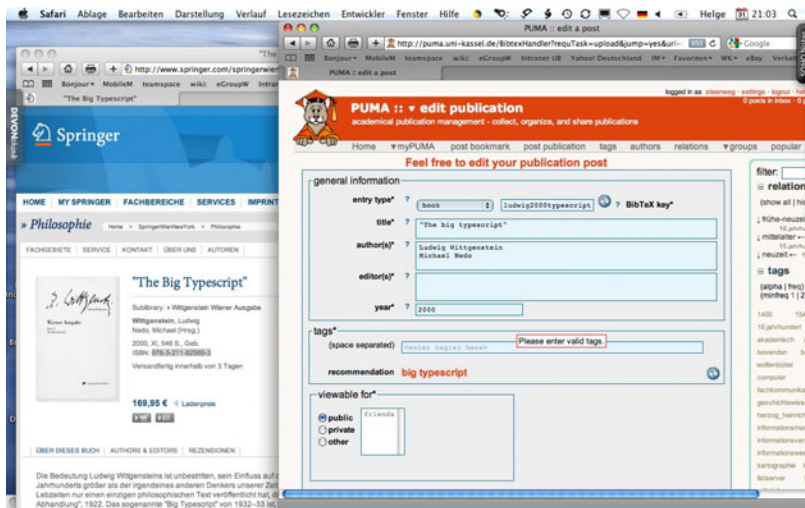


Fig. 2. Using the PUMA-Browser-Plugin to store the publication metadata for an ISBN

Posting a Publication

It is very simple to add publications by entering their metadata manually into the form fields of PUMA. But there are several less laborious ways to get data into PUMA, e.g., file import from BibTeX oder EndNote – or even more conveniently by using a special browser plugin. Every user can then store publication entries while browsing the web. This is accomplished by so-called *scrapers* for over 70 portals (e.g., Amazon, IEEE, Scopus, Muse, BioMed, JSTORE, arXiv, etc.) or library catalogs. Another option is to highlight an ISBN, ISSN or DOI phrase on a web page and to click the postPublication-button of the plugin. This triggers the upload of the associated data to PUMA (s. Fig. 2). All data in PUMA can be edited and exported in over 30 formats. In addition, custom export formats can be created and uploaded by each user.

CV

Another common requirement of authors is to have an easy-to-maintain CV page, featuring all personal details and publications. Within PUMA every user can generate his own CV using a variety of import and export formats. Especially for users of Typo3, a plugin exists with adjustable parameters to maintain an automatically created CV on its own homepage.

5 Next Steps

Currently, we are working on a path to integrate the data automatically into the workflow of an institutional repository, e.g. DSpace. To this end, we implement the SWORD protocol⁶ for easier content exchange between repositories.

Future work also includes better support for research groups (e.g., hierarchical group structures, extended administration options, custom tag sets), reporting functionalities, and a framework for system tags. The improvement of the user interface and the help system are an ongoing effort.

A beta version of PUMA is already available at <http://puma.uni-kassel.de/>. The rollout to all members of the University of Kassel is planned for summer 2010. The system will be made available under an open source licence at the end of the project.

References

1. Benz, D., Hotho, A., Jäschke, R., Krause, B., Mitzlaff, F., Schmitz, C., Stumme, G.: The social bookmark and publication management system BibSonomy. VLDB Journal (to appear)
2. Steenweg, H.: Publikationsmanagement mit PUMA auf der Basis von BibSonomy, <http://www.opus-bayern.de/bib-info/volltexte/2010/865/>

⁶ <http://www.swordapp.org/>

Using Mind Maps to Model Semistructured Documents

Alejandro Bia¹, Rafael Muñoz², and Jaime Gómez³

¹ CIO/DEMI, Miguel Hernández University, Spain
abia@umh.es

² DLSI, University of Alicante, Spain
rafael@dlsi.ua.es

³ DLSI, University of Alicante, Spain
jgomez@dlsi.ua.es

Abstract. We often use UML diagrams for our software development projects, and also for modeling XML DTDs and Schemas (G1), finding that although UML diagrams can effectively be made to represent DTDs and Schemas (either using Class or Component diagrams), in real practice, complex DTDs and Schemas produce unreadable, unmanageable, complex UML diagrams. Recently we started exploring other types of diagrams and unconventional methods which can be both useful for designing and modeling semistructured data, and as teaching aids or thinking tools. This experience also served to open our minds to tools and methods other than the recognized mainstream practices.

In this paper, we describe how we managed to use Mind Maps and a modified Freemind tool to successfully model, design, modify, import and export XML DTDs, XML Schemas (XSD and RNG) and also XML document instances, getting very manageable, easily comprehensible, folding diagrams. In this way, we converted a general purpose mind-mapping tool, into a very powerful tool for XML vocabulary design and simplification (and also for teaching XML markup, or for presentation purposes).

Keywords: Visual Modeling, Mind Maps, XML, DTD, Schema.

1 Introduction

A Mind Map is a tree that develops from a central node, or to say it differently, a set of trees (subtrees) hanging from a central node. A DTD or Schema can be drawn as a graph, with complex interconnections, but if we consider each element and its content model separately, we can draw an element's definition as a tree. Then we can link the contained elements (leaves) to the roots of their corresponding definition trees. This crossed-links will turn the whole diagram into a graph, but with interesting visualization and folding properties. In this way, we can represent a DTD or Schema structure as a set of parallel trees, which closely resemble DTD/Schema syntax, with links connecting some leaves with some roots, in a graph-like manner. The advantage is that trees can be folded and unfolded, allowing us to hide or show different parts of the diagram. This allows for better visualization and comprehension.

1.1 Tool Features

Sometimes, apart from the benefit of using a given type of diagram, tool features are key to a successful visual model. Mind mapping software can be used efficiently to organize large amounts of information, combining spatial organization, dynamic hierarchical structuring and node folding. These features are essential to a good DTD or Schema visual representation.

A good model must be capable of hiding unnecessary detail. This is precisely one of the problems that we had when we used available UML tools with Class diagrams: whole DTD/Schema diagrams were too complex to handle and display, and there was no way to selectively fold parts of them. This, added to the lack of efficient automatic arranging of visual objects while importing a DTD/Schema, made our attempts impractical for real-life purposes.

The ability of Freemind to interactively hide/unhide branches of a Mind Map diagram, and the automatic allocation of nodes around a central point is what makes it so attractive for representing semistructured document structures. User friendly features for copying, pasting, moving, dragging-and-dropping subtrees make it ideal for structure design and editing. For this we needed a way to import and export several types of schemes. So we implemented XSLT transformations for the most popular notations: DTDs, W3C XML Schema and RelaxNG.

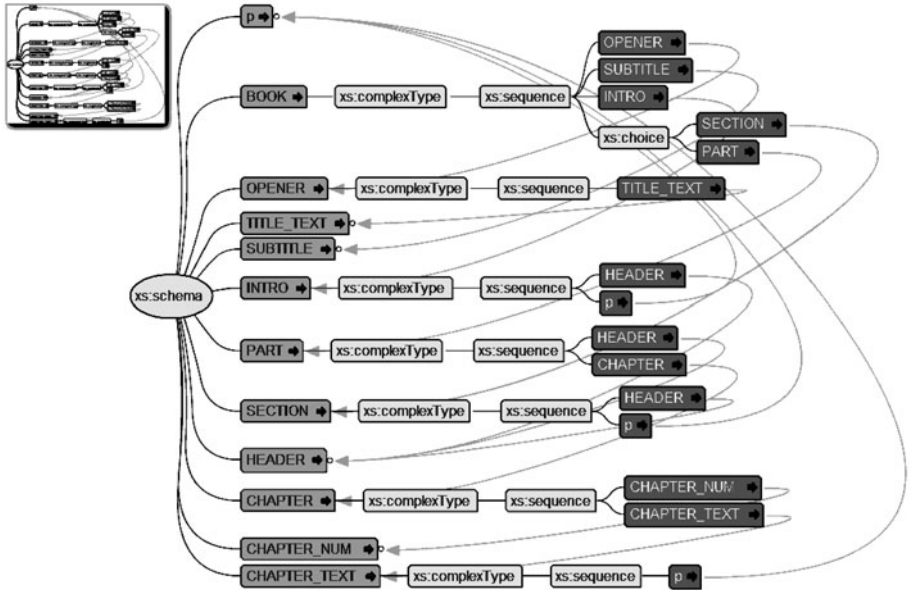


Fig. 1. A simple example of XML Schema represented as a Mind Map, with all the nodes unfolded. The root node ‘BOOK’ is the only node not pointed from any other content model. Note the small map on the top left, which allows seeing the whole picture of a big diagram when it does not fit in the screen.

1.2 Our Implementation

Freemind uses an XML file format to store the diagrams, which is very simple to understand and generate. We have written several XSLT scripts to translate DTDs, XSD and Relax NG Schemas to and from Freemind file format. In the case of DTDs, whose syntax looks quite like, but is not actually XML, we used DTDinst, a program for converting XML DTDs into an XML format. DTDs expressed in this XML format can be easily transformed to any other XML format, and particularly to Freemind Mind Maps. On the opposite direction, we used just an XSLT script in text output mode to write back a DTD.

These XSLT scripts actually build the Mind Map, defining how DTD or Schema elements are rendered. Although the Freemind file format allows assigning an explicit location to graphical elements, it is better not to do so since the program does a very fine job of automatically arranging everything into a nice readable diagram, avoiding overlaps. What we do specify is that nodes be rendered to the left of the initial node. We can also specify whether subtrees should appear folded or unfolded on opening the diagram.

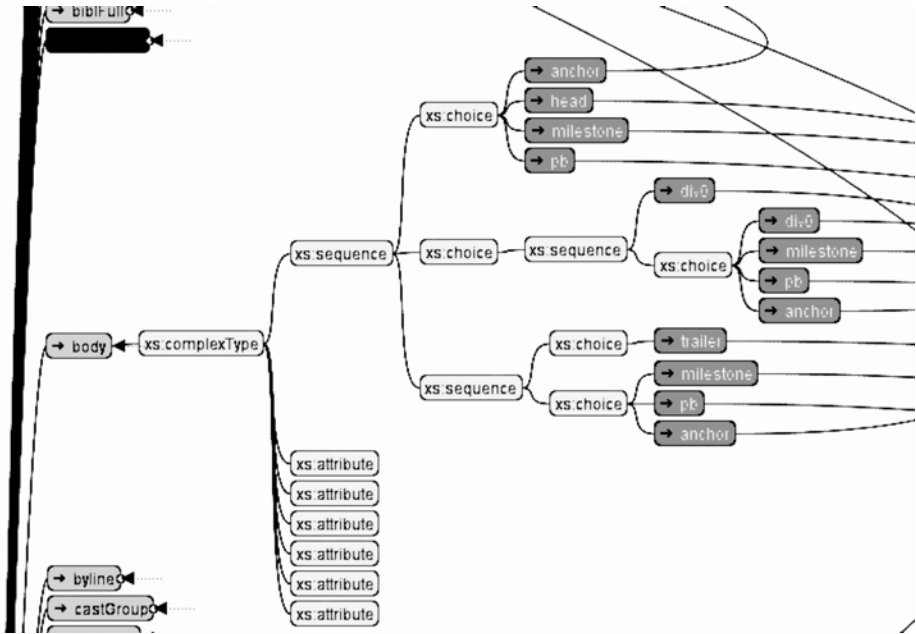


Fig. 2. Partial view of a Mind Map of a TEI XML Schema with the ‘body’ element unfolded, and the rest of the elements folded

The DTD or Schema is presented as a Mind Map beginning in a central oval node, and from it each element definition is a subtree (root in green). Each subtree describes a content model (see figs. 1 and 2), and can be folded and

unfolded by clicking on its green root node. Sequence or choice operators that conform to the rules of the content model are represented in the middle nodes of the subtree (in yellow). Leaf nodes (in red) are the elements referred by the content model, and point (both graphically with curved arrows and by hyperlink) to the subtree definition of the corresponding element (elsewhere in the diagram). This allows for easy navigation and comprehension of the global graph structure. In fig.2, the content model of the element ‘body’ can be clearly seen. The elements referred from this content model (leaf nodes) can be traced by following the curved links, or more easily by clicking on the leaf nodes themselves, which cause the cursor to jump directly to the root of the referred element. In this way, a user can analyze the content model on a unfolded element, and then click on any contained element to directly jump to its root.

In the case of TEI DTDs and Schemas, we have also added external links from each element of the Mind Map to the corresponding TEI documentation page of the element on the Web (see fig. 2). This is very useful for training purposes. It allows the user to immediately see the documentation of any chosen element, just by clicking on it.

1.3 Conclusions and Future Work

We have found many advantages in the use of Mind Maps and the Freemind tool for modeling XML DTDs and Schemas: (1) Automatic organization of graphic elements: key to successful easy import of schemas into a readable diagram. (2) Hierarchical tree structure in an overall graph structure: ideal for representing content models. (3) Information hiding/unhiding by folding tree branches. (4) Powerful visual editing features: copying, pasting, moving, dragging-and-dropping subtrees. There are several tools that can display schemas, but not to edit them in graphical form. (5) Hyperlinking to external files: which we used to automatically link to external documentation pages. (6) Easy XML file format: allowed us to automatically convert DTDs and Schemas to Mind Maps and the other way around, converting the adapted Freemind tool into a document structure design or editing tool.

We have also automated the generation of Mind Map models of XML document instances, and are now working both on DTD/Schema visual comparison, and XML document instance structure visual comparison, to make evident the differences of two of these files by using cross-links and colors on diagrams aligned side by side.

Reference

1. Bia, A., Gómez, J.: UML for Document Modeling: Designing Document Structures for Massive and Systematic Production of XML-based Web Contents. In: Briand, L.C., Williams, C. (eds.) MoDELS 2005. LNCS, vol. 3713, pp. 648–660. Springer, Heidelberg (2005), <http://www.cs.colostate.edu/models05/>

Towards a Public Library Digital Service Taxonomy

Steven Buchanan and David McMenemy

Department of Computer and Information Sciences, University of Strathclyde,
Glasgow, Scotland, United Kingdom

{Steven Buchanan, David McMenemy}@cis.strath.ac.uk

Abstract. Recent research has identified inconsistency of public library digital services, and associated problems of disparity and duplication, as a key usability issue. The hypothesis of this research is that root cause is inconsistent definition and specification of digital services, and that a service taxonomy would facilitate resolution of this issue, providing a classification scheme and controlled vocabulary. Reporting on initial research to validate this hypothesis, which examined options available from 8 of 32 Scottish public library homepages; evidence of inconsistency of terminology and organisation schemes was found, with navigation not always straightforward due to a high number of loosely structured options being available from the majority of sites sampled. Initial findings are discussed including planned second stage research.

Keywords: digital services; usability; service taxonomy; public libraries.

1 Introduction

Within the public domain, digital libraries have the potential to disseminate a wide range of digital content on a wide range of topics, with universal access to digital libraries considered essential to social and economic mobility [1], [2]: however, there is evidence that PL websites have developed in an inconsistent, unstructured, and unplanned manner [3], with significant usability issues reported, particularly relating to navigation and terminology [4], [5], [6], [7]. The central hypothesis of this research is that root cause is inconsistent definition and specification of public library (PL) digital services, and that a service taxonomy would facilitate resolution, providing a classification scheme and controlled vocabulary. This paper reports on stage one research, conducted to establish validity of hypothesis, and to better inform stage two.

2 Digital Services

A digital service is a service or digital resource accessed and/or provided via digital transaction [5]. Services can range from the relatively straightforward, such as provision of online tools and virtual space for collaboration, sharing of content etc., to online reference services, to more complex distributed and interactive systems such as digitized local archive collections purposefully linked to local school curriculum's via virtual learning environments, or cross-institutional integrated digital collections. In

the role of access provider a PL will also establish links to other public information providers with which it shares societal goals such as early learning, cultural heritage, and health and wellbeing.

A service-oriented perspective encourages an organisation to focus on how functions must cooperate to achieve customer satisfaction, transcending the more traditional functional perspective, which can create barriers to information flow and constrain the value that can be generated by the organisation [8]. Further, when formalized and managed (e.g. as a service oriented architecture), services can be repeated, reused, and outsourced.

To the best of our knowledge no taxonomy exists specifically for PL digital services. PL services are defined within literature [9], [10], [11], [12], [13]; however they are not digital library specific, or oriented.

A further challenge is that there is currently exists limited guidance regarding what PL digital services should be, compounded by the nascent state of digital service design and limited previous evaluations of PL Internet services [14], [15]. Further, while digital library evaluation has begun to include service evaluation, it has been largely limited to digital reference services [16] and while usability of digital libraries has been extensively evaluated, usefulness has not, [17] the latter being an aspect of evaluation that might have informed service definition and provision. Accessibility has been considered, but without detailed specification of what exactly should be accessible [18].

3 Research Methodology

In stage one a representative sample of 8 of the 32 Scottish PL websites was sampled (covering a wide ranging demographic, and a major proportion of the Scottish population) to identify and compare the range of options offered from respective homepages. All options from respective homepages were factually recorded, referenced, and listed. Redundant repeat entries were removed (entries sharing exact wording or differing through minor nuance of language but semantically the same). General observations regarding associated aspects of usability were also noted, in particular terminology and navigation, but including aesthetic appearance.

In stage two, and expanding upon stage one sample, content analysis of 32 UK public libraries websites will be conducted (providing a representative national sample). As per stage one, all options from respective homepages will be factually recorded, referenced, and listed, with redundant entries removed. Data will then be disaggregated into meaningful categories through identification of patterns and regularities. Approached inductively, with categories emerging from grouped options available from individual websites (with categories either subdivided or merged with others as appropriate). Anticipated as an iterative cycle of indexing and cross-referencing, shaped and driven by emergent themes and relationships, with the resultant classification scheme presented in hierarchical subject tree format.

Emergent controlled vocabulary guided by established usability principles with regard to terminology, reference to thesauri, and Dewey Decimal Classification scheme and Library of Congress Subject Headings on a case-by-case (subject level) basis if and when applicable.

Field trial of the taxonomy, in particular the controlled vocabulary and identification of preferred and variant terms (assisted by synonym rings).

4 Stage One Findings

202 discreet options (duplicates removed) identified across eight public library homepages. A low rate of reoccurrence found, with no option found repeated on more than four homepages. The minimum number of options displayed was 13, the maximum 69. 5 of the 8 homepages had 29 or more options displayed. When duplicate options were removed minimum remained 13, while maximum reduced from 69 to 59. Significant as more than ten options on the main menu can 'overwhelm' users [16].

A high incidence of ambiguous terminology (27%) and branded terms (20%) found. Again significant, as public perception of terminology is regarded as a key aspect of usability, with simple, natural, and user-oriented language recommended [19]. Technical terms less prevalent (3%).

8 of 8 adopted a hybrid scheme for the organisation of options, variously by topic and task, but extending to user on 7 of 8 homepages (variously providing options by adult, child, migrant, and housebound). 2 of the 8 used inconsistent terminology for (repeated) options on the homepage. Navigation considered problematic due to the high number of options available, and limited or ambiguous associated organisation schemes.

5 Conclusion

Stage one research has identified a total of 202 discreet options available across eight PL homepages, with the majority providing 29 or more options under limited or ambiguous organisation schemes. A low rate of reoccurrence found. A high incidence of ambiguous terminology (27%) and branded terms (20%) found, and navigation considered problematic due to the high number of options available and limited or ambiguous associated organisation schemes.

Findings would appear to support the hypothesis, that a root cause of PL website usability issues (and disparity and duplication) is inconsistent definition and specification of PL digital services, and that a service taxonomy would facilitate resolution of this problem.

References

1. Liu, Y.Q., Martin, C., Roehl, E., Yi, Z., Ward, S.: Digital information access in urban/suburban communities. *OCLC Systems & Services: International Digital Library Perspectives* 22(2), 132–144 (2006)
2. European Commission Information Society and Media. i2010: Digital Libraries (2006), http://ec.europa.eu/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf
3. Brinkley, M.: The future of library websites. *VINE* 113, 18–25 (2007)

4. Atherton, L.: seamlessUK – building bridges between information islands. *New Library World* 103, 467–473 (2002)
5. Socitm. Better connected 2010: a snapshot of all local authority websites (2010), http://www.socitm.net/downloads/file/506/better_connected_2010-full_report
6. McMenemy, D.: Internet identity and public libraries: communicating service values through web presence. *Library Review* 56(8), 653–657 (2007)
7. Harden, S., Harden, R.: Why are we waiting? Observations on how UK public libraries are using the world wide web. *VINE* 113, 8–12 (2007)
8. Gibb, F., Buchanan, S., Shah, S.: An integrated approach to process and service management. *International Journal of Information Management* (26), 44–58 (2006)
9. Dempsey, L.: Scientific, Industrial and Cultural Heritage: a shared approach: a research framework for digital libraries, museums and archives. *Ariadne* 22 (2000)
10. Brophy, P.: *The Library in the Twenty-First Century: new services for the information age*. Library Association, London (2001)
11. Dewe, M.: *Planning Public Library Buildings: concepts and issues for the librarian*. Ashgate, Aldershot (2006)
12. Chowdhury, G.G., Burton, P.F., McMenemy, D., Poulter, A.: *Librarianship: an introduction*. Facet, London (2008)
13. McMenemy, D.: *The Public Library*. Facet, London (2008)
14. Williams, K., Chatterjee, S., Rossi, M.: Design of emerging digital services: a taxonomy. *European Journal of Information Systems* 17, 505–517 (2008)
15. Aitta, M.J., Kaleva, S., Kortelainen, T.: Heuristic evaluation applied to library web service. *New Library World* 109(1/2), 25–45 (2007)
16. Xie, H.: Evaluation of digital libraries: criteria and problems from users' perspectives. *Library & Information Science Research* 28(3), 433–452 (2006)
17. Buchanan, S., Salako, A.: Evaluating the usability and usefulness of a digital library. *Library Review* 58(9), 638–651 (2009)
18. Ghauri, P., Gronhaug, K.: *Research methods in business studies*. FT Prentice Hall, Essex (2005)
19. Molich, R., Nielsen, J.: Improving a human-computer dialogue: What designers know about traditional interface design. *Communications of the ACM* 33(3) (March 1990)

Multimodal Image Collection Visualization Using Non-negative Matrix Factorization

Jorge E. Camargo, Juan C. Caicedo, and Fabio A. González

Bioingenium Research Group
National University of Colombia
{jecamargom, jccaicedoru, fagonzalezo}@unal.edu.co

Abstract. In this paper we address the problem of generating an image collection visualization in which images and text can be projected together. Given a collection of images with attached text annotations, we aim to find a common representation for both information sources to model latent correlations among the collection. Using the proposed latent representation, an image collection visualization is built, in which images and text can be projected simultaneously. The resulting image visualization allows to identify the relationships between images and text terms, allowing to understand the semantic structure of the collection.

1 Introduction

Image collection exploration has been shown to be a promising strategy for image retrieval [1]. In this strategy, the user interacts with the system while it learns from the user's feedback to deliver more precise results. Image collection visualization plays an important role in this process. To construct the visualization, an image is represented in a two dimensional space, in which images with similar features are mapped to neighboring positions. In this way, it is expected that users can access images with similar properties in the same region of the screen. This mapping allows the user to see image inter-relationships and to easily identify how they are semantically related.

The visualization may be built using the visual content of the image to organize the image collection according to some visual similarity properties rather than meaningful semantic criteria. Thus, images with similar low-level visual features may appear in the same region of the screen even though they represent different semantic concepts. To provide a more semantic and organized visualization, some learning approaches have been proposed to adapt the position of images in the screen according to the user's preferences [2] or some predefined semantic categories in the collection [3]. However, these schemes have two limitations that prevent them from being used for analyzing massive image collections. First, they exclusively project only visual information. Second, these learning algorithms rely on user's feedback or structured metadata to learn the semantic organization of images, requiring expensive efforts to collect user profiles or reliable annotations.

In this paper we use Non-negative Matrix Factorization (NMF) to construct a multimodal latent space in which visual features and text terms are represented together. The location of text terms and visual features can be identified in the latent space, bringing a unified way to analyze the relationships between both modalities. This latent multimodal representation is then used to project both, text terms and images in the same 2D plane to construct an image collection visualization with a semantic organization given by text data and at the same time marks the regions of the screen in which semantic concepts can be found.

2 Multimodal Image Collection Visualization

The image database is composed of two data modalities, herein denoted by X_v and X_t . The former is a matrix whose rows are indexed by n visual features and whose columns are indexed by l images. The latter has m rows to represent text terms and l columns for images as well. The construction of a latent semantic space is based on the simultaneous analysis of visual features and text terms to generate a semantic space for image indexing, by exploiting multimodal relationships. The proposed strategy uses a multimodal matrix $X = [X_v^T X_t^T]^T$ that is factorized using NMF as $X_{(n+m) \times l} = W_{(n+m) \times r} H_{r \times l}$, where W is the basis of a latent space in which each multimodal object is represented by a linear combination of the r columns of W . The corresponding coefficients of the combination are codified in the columns of H .

We find this factorization using NMF based on the divergence objective function as is described in [4]. One important aspect in the proposed multimodal scheme is that we can represent both text terms and images in the same space. The latent semantic space is indexed by r latent factors. Each image is represented by r values that can be understood as the membership degree of each image to each of r clusters. In the same way, each of the m text terms have a representation in the latent semantic space given by the rows of the matrix $W_{m \times r}^t$. Thus, since the position of text terms is known, we can analyze their neighborhood to understand image semantics.

Finally, to generate the image collection visualization we use Principal Component Analysis (PCA) to project text data and images in a 2-dimensional coordinates system, taking their representation in the latent space. As input, PCA receives a representation matrix $T = [W_{rxm}^T H_{rxl}]$, where W_{rxm}^T is the representation of the m text terms in the latent space and H_{rxl} is the representation of all images in the latent space.

3 Experimental Evaluation

In this evaluation, we used a subset of the Corel image database which is composed of 2,500 images in 25 categories. Visual image content is represented using a bag of features approach: each image is split in non-overlapping blocks of 8×8 pixels, and for each block the SIFT descriptor is computed. Then, using an image training set, a codebook of 1,000 blocks is built using the k -means algorithm.

The matrix X_v^T is constructed using a vector in \mathbb{R}^{1000} for each image. The names of the categories were used as text terms. To build X_t^T , a binary vector in \mathbb{R}^{25} for each image is built using the category information, in which the i -th position is 1 if the image belongs to i -th class and 0 otherwise. This information is used to simulate keywords associated to image contents following a bag of words approach for text data. Finally, we computed the NMF factorization as $X_{(1000+25) \times 2500} = W_{(1000+25) \times 30} H_{30 \times 2500}$. We set the value of the r parameter empirically to 30, which was determined by maximizing a standard measure (mean average precision) in an image retrieval task. The concatenation of both feature vectors for each image was normalized to have norm $\ell_2 = 1$.



Fig. 1. Multimodal visualization of the complete dataset

A multimodal visualization of the complete data set is illustrated in Figure 1. Even though some images are occluded and the layout has not been optimized, the user can get oriented in the metaphor thanks to the presence of the text terms in the visualization. It is especially remarkable that some similar text terms, from the semantic viewpoint, are represented closely in the latent space, and therefore in the visualization as well. For instance, notice the close position of the terms *plants* and *forest* as well as *beach*, *boats* and *isles*. Individually, all of them are identified as completely different categories, but they share many similarities in terms of visual properties, as well as from the semantic perspective. The coherence of their positions supports the idea that the NMF algorithm is

providing a consistent representation for images since a semantic perspective, and also shows how the visual patterns are revealing meaningful connections between text terms. Other examples can be found by observing the positions between *volcano-mountain* and *flags-cards*.

Also, the quadrants of the visualization can be conceptualized with some other high-level interpretations of the image categories. For instance, the concepts at the bottom part of the visualization may be associated to open landscapes, such as those for *boats*, *beach*, *aviation* and *mountains*, among others, while the upper part, may be associated to closed landscapes for *roses*, *fruits*, *cats* and *dogs*. The left region may be associated to more artificial scenes such as those related to *cats*, *flags*, *cards* and *drinks*, while the right region may be associated to natural landscapes. This indicates that the combined latent semantic representation is in fact capturing some aspects of the true semantics of the collection.

4 Conclusions and Future Work

This paper has presented a first step towards the construction of a semantic image collection exploration system that allows to understand the distribution of images in the collection. To bring a semantic organization of the image collection, we propose the joint analysis of image features and text terms to construct a meaningful visualization.

We used a Non-negative Matrix Factorization algorithm to built a latent space for multimodal data, in which images and text terms can be represented together. We showed the potential of the proposed strategy, following a qualitative evaluation of the resulting collection visualizations. The first clear advantage of the proposed approach is the ability to locate text terms in the 2D canvas to guide the user in a hypothetical exploration process. This result makes an important difference among the state-of-the-art methods for image collection exploration, that are mainly based on visual features without a clear identification of the regions in the screen.

References

1. Heesch, D.: A survey of browsing models for content based image retrieval. *Multimedia Tools and Applications* 40(2), 261–284 (2008)
2. Bartolini, I., Ciaccia, P., Patella, M.: Pibe: Manage your images the way you want! In: *International Conference on Data Engineering*, pp. 1519–1520 (2007)
3. Ding, H., Liu, J., Lu, H.: Hierarchical clustering-based navigation of image search results. In: *MM 2008: Proceeding of the 16th ACM International Conference on Multimedia*, pp. 741–744. ACM, New York (2008)
4. Lee, D.D., Seung, H.S.: Algorithms for nonnegative matrix factorization. *Advances in Neural Information Processing Systems* 13, 556–562 (2001)

A New Perspective on Collection Selection

Helen Dodd¹, George Buchanan², and Matt Jones¹

¹ Department of Computer Science, Swansea University, United Kingdom

² Centre for HCI Design, City University, London, United Kingdom

Abstract. Collection selection is traditionally a sub-problem of meta-search, and identifies collections most likely to contain relevant documents. However, we propose to treat collection selection as an independent search task with the goal of identifying collections that are relevant as a whole; so the user may return to them to serve future (related) information needs. Using a new methodology and framework we evaluate the suitability of existing collection selection algorithms for this search task, compared with a new algorithm designed specifically for the task.

Keywords: Collection selection, database selection, collection ranking.

1 Introduction

Consider a scenario where a user wants to locate authoritative collections (e.g. digital libraries) on a particular topic, to fulfil both current and *future* information needs. A technique to identify collections with a degree of relevance to a query is *collection selection*; a sub-problem of metasearch. It supports *document retrieval* by choosing a subset of collections most likely to contain relevant documents. The query is dispatched to these collections, and the results merged to form a list of relevant documents [5]. We propose to treat collection selection as an independent search task. Here the goal is not to find individual documents from multiple collections, but to identify individual collections containing a high *proportion* and *quantity* of relevant documents: collections *about* the query.

We present a methodology and framework for the evaluation of algorithms over our interpretation of collection selection. These techniques are used to evaluate the performance of a new algorithm, designed specifically for our task; and to examine the suitability of existing collection selection algorithms.

2 Our Evaluation Method

To test the suitability and performance of algorithms for our retrieval task we use *scenario* and *optimal performance* tests. The scenario tests use controlled data to scrutinise the behaviour of algorithms over clear cases: algorithms producing incorrect rankings will not suit our needs. Each of our seven scenarios models three collections, one of which is the clear winner, another the clear loser. Different attributes (size, term frequency, quantity/proportion of relevant documents) of the collections are varied in each scenario.

The optimal performance test surveys how well algorithms estimate¹ an *optimal* ranking. Traditionally [2] the optimal orders collections by the number of relevant documents they contain. However, we propose an optimal more representative of our search task; where suitable collections contain a high number and proportion of relevant documents. We represent this with two metrics:

$$RS_c = \frac{|\text{relevant documents in collection}|}{|\text{relevant documents}|} \quad RP_c = \frac{|\text{relevant documents in collection}|}{|\text{documents in collection}|}$$

where RS_c is the *share* and RP_c the *proportion* of relevant documents in a collection c . We order collections by decreasing harmonic mean (F-score) of RS_c and RP_c : the F-score Based Ranking (FsBR). We use the Spearman rank correlation coefficient to determine how well algorithm rankings estimate the optimal.

3 Our Algorithm

Our algorithm (called *Doddle*) is inspired by criteria for highly ranked collections [8]: if each query term is *common* and occurs *frequently* in a high *proportion* of documents within a collection (relative to other collections), the collection should be highly ranked. Doddle ranks collections in decreasing order of *merit*. For a given query q , the merit associated with collection c is calculated by:

$$\text{merit}(q, c) = \sum_{t \in q} f_{q,t} \times (RC_{t,c} + RP_{t,c} + RF_{t,c})$$

where $f_{q,t}$ is the number of occurrences of term t in the query and:

$$RC_{t,c} = \frac{C_{t,c}}{\sum_{i=1}^{|C|} C_{t,i}} \quad RP_{t,c} = \frac{P_{t,c}}{\sum_{i=1}^{|C|} P_{t,i}} \quad RF_{t,c} = \frac{F_{t,c}}{\sum_{i=1}^{|C|} F_{t,i}}$$

(Relative Commonness) (Relative Proportion) (Relative Frequency)

where:

- $C_{t,c}$ = $\frac{f_{c,t}}{\text{tokens}_c}$ (commonness of term t in collection c);
- $P_{t,c}$ = $\frac{df_{c,t}}{\text{docs}_c}$ (proportion of documents in collection c containing term t);
- $F_{t,c}$ = $\frac{f_{c,t}}{df_{c,t}}$ (average occurrences of term t in documents in collection c);
- $f_{c,t}$ is the number of occurrences of term t in collection c ;
- tokens_c is the total number of terms in collection c ;
- $df_{c,t}$ is the number of documents in collection c containing term t ; and
- docs_c is the total number of documents in collection c .

4 Apparatus

We support the evaluation of algorithms with a set of applications that enable the management of collection data, creation of scenarios, and the execution of tests and the display of their results.

¹ Methods indicating how well an algorithm estimates an optimal include: Mean-squared error; Spearman rank correlation coefficient; and a recall-based metric [2].

Our *optimal performance tests* use 16 collections, ranging from 16 to 800,000 documents in size. Their coverage is varied: some specialise in one subject, others address a range of subjects. They are real collections, harvested using OAI-PMH². We harvest the Title and Description fields in Dublin Core format and create two indexes from the data: “title only” and “title and description”.

Previous studies have created artificial collections from TREC data (divided by source and date, or size [6]). Such collections are often of generalist material, and are thus a poor substitute for those we aim to serve: specialised and of varying size. However, using real collections leaves us without document relevance judgements, required by FsBR in the optimal performance tests. We therefore generate sudo-relevance judgements using document ranking algorithms. The Apache Lucene³ library is used to create a document index from the harvested metadata. For each query, the documents are ranked by *tf.idf*, BM25 and the Lucene search algorithm. Documents that all three algorithms agree are relevant are appended to a list of “relevant” documents. From this we determine the number of relevant documents in each collection, and calculate their F-scores.

We use a test set of 50 queries (1-10 terms long). Some queries target specific collections, others describe wide subject areas, present in multiple collections.

5 Experimental Results

Our initial experiments use our scenario and optimal performance tests to survey the suitability of existing algorithms for our task, and compare their performance to Doodle. We investigate algorithms previously shown to be effective: CORI [1] (often used as a benchmark); bGLOSS [3]; Cue Validity Variance (CVV) [7]; and Inner Product [8]. We also consider Average Inverse Collection Term Frequency (AvICTF) [4], a *query performance predictor* which will produce rankings based on the predicted quality of results (were each collection searched in isolation).

In the scenario tests, Inner Product, CVV and AvICTF were found to be ill-suited to our search task; failing on one, two and five scenarios respectively. Specifically, AvICTF frequently favoured collections with the least query term occurrences. CORI and bGLOSS succeeded for all seven scenarios, suggesting they may be suitable baselines for comparison with our own algorithm; which also produced correct rankings in all scenarios.

Table 1 shows the average correlations between the algorithm rankings and the optimal. We also compare the effect of using only title metadata, versus both titles and descriptions. Our algorithm produces rankings with a much higher correlation to the optimal ranking than any of the existing algorithms. However, for an average of eight collection results per query, the correlation is below the 5% significance level: there is still considerable room for improvement.

Most algorithms produced a better correlation with the optimal when queries were executed over the “title only” term index. One explanation may be that the description metadata has more noise, however further investigation is required.

² <http://openarchives.org/OAI/openarchivesprotocol.html>

³ <http://lucene.apache.org/>

Table 1. Average Spearman rank correlation coefficients for each algorithm

Algorithms	Titles	Titles and Descriptions
AvICTF	-0.537	-0.684
CVV	-0.035	-0.179
Inner Product	0.037	0.007
bGLOSS	0.149	0.153
CORI	0.226	0.106
Doddle	0.624	0.518

6 Conclusions and Future Work

We have presented a new interpretation of collection selection: to treat it as an independent search task, with the goal of identifying quality collections that will satisfy a user's current and future information needs. The paper reported our methodology to evaluate algorithms for this task. With this, we investigated the performance of existing collection selection algorithms, compared to that of our own algorithm. Our algorithm significantly outperformed existing algorithms; however, its correlation with an optimal ranking was still not satisfactory.

Our future work will iteratively improve our evaluation methodology and the selection algorithm. This will include the refinement of the scenario and optimal performance tests. In particular, we aim to ensure our optimal ranking produces the most sensible ordering of collections. Future experiments will evaluate algorithms in terms of a baseline ranking, and use additional evaluation metrics.

Acknowledgements. Helen Dodd is supported by an EPSRC Doctoral Training Grant.

References

1. Callan, J.P., Lu, Z., Croft, W.B.: Searching distributed collections with inference networks. In: Proc. SIGIR, pp. 21–28. ACM Press, New York (1995)
2. French, J.C., Powell, A.L.: Metrics for evaluating database selection techniques. *World Wide Web* 3(3), 153–163 (2000)
3. Gravano, L., García-Molina, H., Tomasic, A.: The effectiveness of GLOSS for the text database discovery problem. In: Proc. SIGMOD, pp. 126–137. ACM Press, New York (1994)
4. He, B., Ounis, I.: Query performance prediction. *Inf. Syst.* 31(7), 585–594 (2006)
5. Meng, W., Yu, C., Liu, K.L.: Building efficient and effective metasearch engines. *ACM Comput. Surv.* 34(1), 48–89 (2002)
6. Powell, A.L., French, J.C.: Comparing the performance of collection selection algorithms. *ACM Trans. Inf. Syst.* 21(4), 412–456 (2003)
7. Yuwono, B., Lee, D.L.: Server ranking for distributed text retrieval systems on the internet. In: Proc. DASFAA, pp. 41–50. World Scientific Press, Singapore (1997)
8. Zobel, J.: Collection selection via lexicon inspection. In: Proc. ADCS, pp. 74–80 (1997)

Creating a Flexible Preservation Infrastructure for Electronic Records

Karen Estlund and Heather Briston

University of Oregon
{kestlund,hbriston}@uoregon.edu

Abstract. As universities begin to address their first significant collections of electronic records, the needs of the collections often outstrip the resources and support available. This poster will illustrate the steps taken to transition and preserve a presidential electronic records collection into an university archives with limited systems support and preparation for future preservation needs. The infrastructure created was designed to quickly ingest at-risk records and allow for file migration and system evolution as future technologies are implemented.

Keywords: Digital Preservation, Digital Libraries, Preservation Planning, Institutional Archives, Migration.

1 Introduction

The transition of a presidency is always a momentous occasion and that is certainly the case in a public university. In summer 2009, the University of Oregon president retired after fifteen years. Although previous presidents had sporadically created electronic records, this was the first presidency to create significant amounts of electronic records along with traditional paper records. The level of decisions documented and the breadth of topics covered make the records of the presidency the most important collected by the University Archives and Libraries. Under Oregon Administrative Rule 166-475, the Oregon University System records retention schedule, [1] the bulk of records created in the office are deemed as permanent and must be transferred to the archives when no longer active. This collection of records created a sense of urgency to collect the records before any loss, a need to integrate discovery with the accompanying paper materials, and a need for an infrastructure in the Libraries. With this new collection, an opportunity arose to create a new organizational collaboration between the university historian and archivist, the digital collections coordinator, and library systems personnel, creating new working standards for this type of collection. Without an ideal out of the box system to implement, the collaboration concentrated on creating a standards compliant infrastructure where records can easily be migrated into a system in the future.

2 Preservation Planning

Ideally, a repository system based on the Open Archival Information Standard (OAIS) [2] would have been implemented for this collection. The two digital asset management

systems used by the University Libraries, CONTENTdm and DSpace, did not fully meet the collection's needs, which needed to be described on a collection rather than item level and did not require a web presence. The presidential electronic records contain over 6,000 files and additional embedded files. In addition, this collection was unique in the need to fully integrate retrieval of the paper documents along with the electronic records and not support separate systems for each format of electronic files (e.g. images vs. email). In order to facilitate a system that could quickly be constructed for secure ingest, we attempted to follow the principles of the OAIS model with manual controls in an infrastructure where selections for automations and migration to an OAIS compliant repository can easily be added. We used the PLATTER documentation for planning for a trusted repository [3] to help guide the decisions for ingest, migration schedules, institutional support, and access.

3 Administration

The president's office was accustomed to providing paper records to University Archives; however, although they are ubiquitous today, even the president's office does not think about the long term preservation and access of electronic records. There is a level of trepidation when it comes to the transfer of electronic records to the archives because of the sensitive nature of materials. While most documents created by the university are public records and subject to the Oregon Public Records law [4], there are numerous exemptions from disclosure, as well as other state and federal laws that require documents or information to be kept confidential. Many records creators at the university are concerned about the ease of inadvertent disclosure of electronic records, especially outside of their context.

Once transfer was agreed, we worked with the Executive Assistant to the President to prepare the electronic records. She went through the documents and e-mail accounts and filed many messages, discarded junk mail, as well as flagged confidential or otherwise sensitive items prior to our ingest of the records. In a collaborative meeting with campus IT and the president's office, we paved the way to insure everyone was comfortable with the records transfer, security and access. In this situation, the total transfer was less than 4GB, and for expediency was transferred via a DVD. In the immediate future, the Libraries will utilize Windows file sharing and active directory, so that files can be moved and ingested without transfer via media.

Management of permanent electronic records is a long term, labor intensive commitment. The total effort involved includes not only the staff and activities involved in the initial transfer, but an ongoing commitment by office contacts, university archives staff, the libraries' digital collections coordinator, and the libraries' systems staff.

4 Ingest

The preparation conducted by the former Executive Assistant to the President allowed for all permanent electronic materials to be transferred in one batch including: the official university records from the president and his assistant and the president's

personal records. The first step in turning these personally managed collections into a library archives collection required an inventory of the file types found among the records. During the evaluation of the files, we also took steps to prepare the files for transfer including changing file names to standard forms without special characters or spaces and ensuring that proper file extension were applied to all files.

We created a plan for migration for files in proprietary and unsustainable formats. The native files and the converted file type are stored in the preservation copy of the collection. A majority of the office documents were converted to PDF files, which we were able to utilize batch processing migration tools. Since .pst file format has recently been released as an open standard, we opted to contain the email in the original format and put on the list of file types to monitor for future migration.

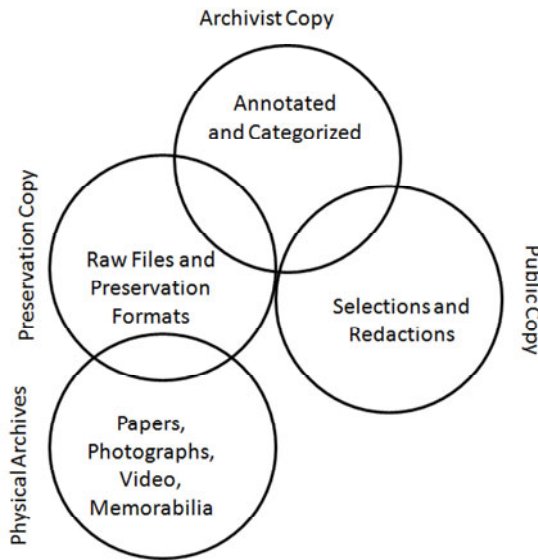


Fig. 1. Preservation and Access of Presidential Records

5 Archival Storage, Data Management and Access

After taking the multiple personal information management systems and combining them into one record of an individual and his job as president of the university, we created three distinct areas for the collection:

1. The preservation layer consists of the files in their native format and structure delivered by the President's office along with migrated versions of the files. Records that were created in Archivists' Toolkit¹ to describe the files are exported into EAD XML and live on the server with the archival files. The files are backed up in multiple locations and check-sums are run to avoid bit rot. The log of the file

¹ Archivists' Toolkit, <http://www.archiviststoolkit.org/>

types created during the inventorying process are kept to help monitor future migration needs if they arise. Because issues of confidentiality, privacy, and state and federal record laws apply, access to this section is restricted.

2. The archivists' layer consists of the preservation format of the files and is organized and tagged according to the system devised by the university historian and archivist. Records in Archivists' Toolkit are used to describe the content and point to the server location of these files along with the paper records.
3. The public access layer is a redacted copy where files have been determined not to breach confidentiality or contradict any laws guiding access. These files are available on a file server that allows for designated public terminals and staff computers to access read-only versions of the files. Future plans include providing access to files online, as risk is assessed, and integrating into the existing UO Office of the President's Digital Collection.²

6 Conclusion

Although it was unrealistic to implement a fully OAIS compliant repository in time to collect these important records, by following the tools and standards provided by the OAIS model and the PLATTER toolkit, we were able to implement a transitional system that meets current needs and can easily be adapted as future technologies are integrated. The success of the system will be measured by the ability to preserve digital objects and retrieve relevant records from paper and electronic collections.

References

1. Oregon Administrative Rule, Secretary of State, Archives Division, Oregon University System Records, <http://arcweb.sos.state.or.us/>
2. Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1 Blue Book (January 2002), <http://public.ccsds.org/publications/archive/650x0b1.pdf>
3. DigitalPreservationEurope, DPE Repository Planning Checklist and Guidance DPED3.2 (April 2008), http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf
4. Oregon Revised Statutes, ch. 192, Records, Public Reports and Meetings (Public Records Law), <http://www.leg.state.or.us/ors/192.html>

² <http://oregondigital.org/digcol/uopres/>

Matching Intellectual Works for Rights Management in the European Library

Nuno Freire

The European Library, The National Library of the Netherlands,
Willem-Alexanderhof 5, 2509 LK The Hague, Netherlands
nfreire@gmail.com

Abstract. This poster presents the work matching system implemented in The European Library for identifying different publications with the same underlying intellectual work. This work is contextualized in the rights management framework of project ARROW, where The European Library is the main source of bibliographic metadata as an aggregator of Europe's national library catalogues.

Keywords: copyright, entity matching, intellectual work, bibliographic metadata.

1 Introduction

In Europe, high-level discussions have taken place among the relevant stakeholders, about how to maximize access to digital content. These include libraries, publishers and collective rights organisations.

Libraries hold materials of public interest, which they may be willing to digitise and make public, but they need to know the copyright status of those works. A major challenge that must be overcome is the significant fragmentation of the rights information infrastructure that exists at present, since it makes the copyright clearance process very demanding and expensive for libraries.

In project ARROW¹ (Accessible Registries of Rights Information and Orphan Works) a single framework is being established to manage rights information. It proposes to create a seamless service across a distributed network of national databases containing information that will assist in determining the rights status of works. This infrastructure, once established², will provide valuable tools for libraries and other organisations to contact rights holders in seeking copyright clearance for the use of content.

This poster presents the work matching system implemented in The European Library³ for identifying different publications with the same underlying intellectual work. This system supports rights management framework of project ARROW, where The European Library is the main source of bibliographic metadata as an aggregator of Europe's national library catalogues.

¹ <http://www.arrow-net.eu> (this project is funded under the eContentplus programme).

² The first release of the ARROW system prototype is scheduled for May 2010.

³ <http://www.theeuropeanlibrary.org>

2 The ARROW Workflow

A library wishing to digitise a book has to go through a number of steps:

- To identify the underlying work incorporated in the book to be digitised
- To find out if the underlying work is in the public domain or in copyright, an orphan work or out-of-print
 - To describe clearly the use that is requested for the book, such as digitisation for preservation, electronic document delivery etc.
 - To identify the rights holder(s) or their agent, such as a collecting society
 - To seek the appropriate permission

ARROW addresses the interoperability of rights information along this process. It must support the identification of a work, the clarification of its rights status and the identification of the rights holders.

This infrastructure depends on the availability of existing bibliographic data and rights information. There is already an established and generally well-regarded information infrastructure for print material, through national bibliographies, books in print and the databases of rights organisations. Currently, these sources are not interoperable because of differences in data collection policies and data schemas. Bibliographic databases rarely include metadata about rights ownership and usage policies. Such information is usually held in a wide array of formats by publishers, collecting societies and authors.

Bibliographic data from the catalogues of Europe's national libraries are one of the key data sources in ARROW. ARROW is therefore building on the existing interoperability achieved through The European Library. Launched as an operational service in March 2005, The European Library is a free service that offers a single point of access to the bibliographical and digital collections of the National Libraries of Europe. 46 of the 48 national libraries in Europe have included their collections in The European Library. This resource is being used as a core source of bibliographic data within ARROW's infrastructure.

Figure 1 shows an overview of the basic workflow of ARROW. It starts from a library as a potential user that wishes to digitise a book and shows the process that the ARROW system must support to provide a response containing the requested rights information. The process depends on data from several sources:

- Library bibliographic data aggregated in The European Library (TEL)
- Author data from the Virtual Authority File (VIAF)
- Publishing data from Books In Print database (BIP)
- Rights holders data from Reprographic Rights Organizations (RRO)
- The International Standard Text Code (ISTC), a numbering system developed to enable the unique identification of textual works.

The initial steps of the workflow depend on The European Library's central index for two tasks: to identify the exact record of the book that the library wants to digitise; and to identify other records of books that share the same underlying intellectual work. All records that share the same underlying work form a cluster, which will be used in the following task of the workflow to identify the rights information required by the library.

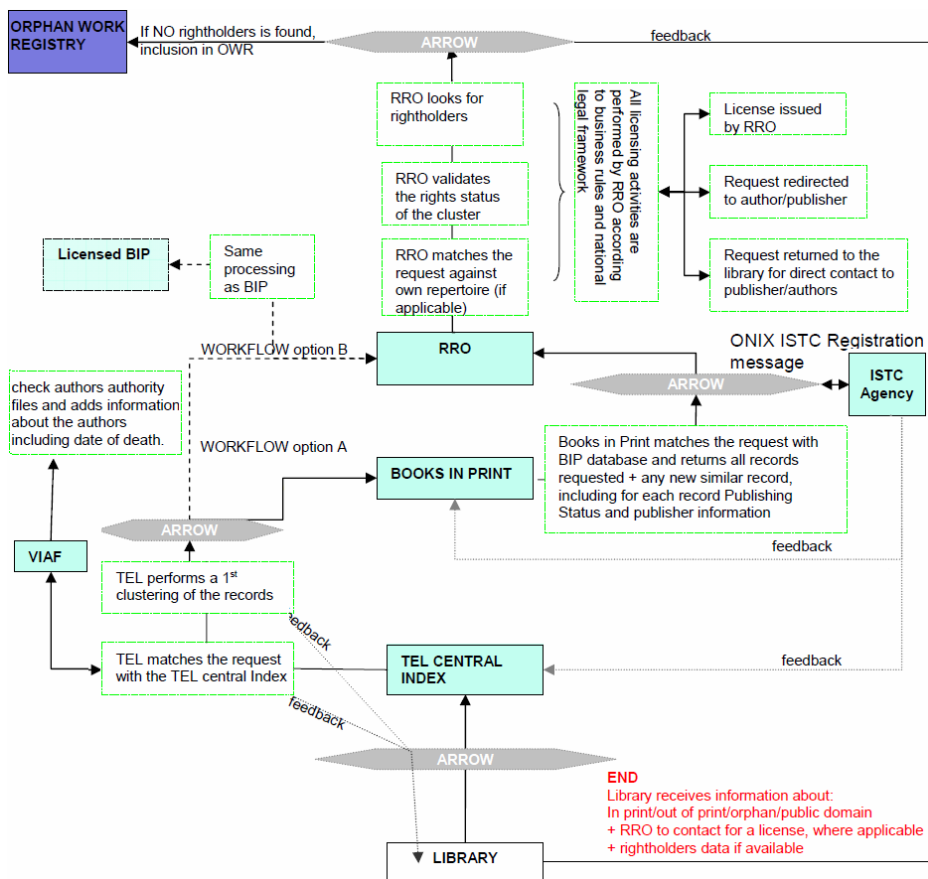


Fig. 1. The ARROW workflow

3 Extraction and Matching of Work Metadata

The European Library’s central index contains bibliographic records in different MARC formats (in the course of ARROW two formats will be addressed: MARC21, which will be used as the primary metadata interoperability format for ARROW and UNIMARC, which is used in a wide number of national libraries across Europe). These records describe manifestations and typically do not explicitly or identify the underlying intellectual work.

In order to fulfil the requirements of ARROW, a work matching system was built to provide The European Library central repository with the capability to extract the underlying work from the manifestation records and match them with those of other records. This system is represented in Figure 2.

The work matching system was built as an ETL (Extraction, Transformation and Loading) process, a typical approach for building data warehouses. It starts with the preparation of data for further processing. This step comprises tasks for selecting the relevant data from the bibliographic records, parsing it when necessary, and doing

initial transformations so that the underlying work is represented according to a standardized work schema.

Two standard options for representing intellectual works were evaluated: the Functional Requirements for Bibliographic Records (FRBR) and the “work metadata” of the ISTC. The final decision was to use ISTC since the concept of “work” in ISTC is closer to the objectives of ARROW than that of FRBR, which is focused on library users needs, making it less suitable for copyright clearance purposes.

To match the references to works, it was necessary to compare the values using string similarity metrics, since variations in the way data is encoded during cataloguing are frequent (typing errors, misspellings, different practices, etc), and because the cost of missing a match may be very high when dealing with copyright.

To avoid having a comparison algorithm that would require comparing all records with all other records, that is, having quadratic complexity, n-grams of size 4 of the titles and authors of the works were indexed using Lucene. Only those records with similar n-grams are compared. This allowed reducing the number of record comparisons to an acceptable number, which scales to the size of the central repository, allowing requests, received from ARROW, to be answered in seconds.

An interface to the work matching system is provided for the central ARROW system by means of web services.

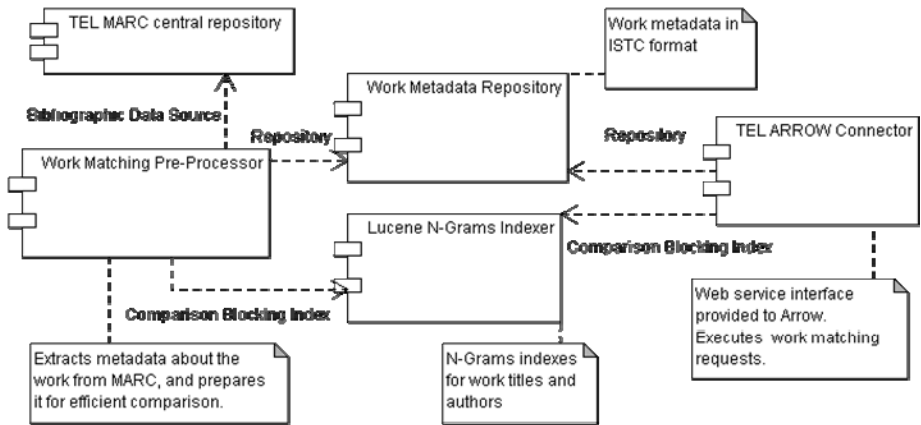


Fig. 2. The European Library Work Matching System for ARROW

4 Conclusion

ARROW supports the European Commission’s i2010 Digital Library vision by providing the means to clarify the rights status of works, allowing its digitization and inclusion within Europe’s digital libraries.

The work matching system of The European Library provides ARROW with the capacity to identify the common underlying work of several publications, a vital part of the rights clearance process.

Work matching will have future application in The European Library. It is envisaged that this system may be applied for other purposes, mainly to improve The European Library’s portal functionality for end users with FRBR work matching.

Mopseus – A Digital Library Management System Focused on Preservation

Dimitris Gavrilis¹, Christos Papatheodorou^{1,2},
Panos Constantopoulos^{1,3}, and Stavros Angelis¹

¹ Digital Curation Unit – IMIS, Athena Research Centre, Athens, Greece

² Department of Archives and Library Sciences, Ionian University, Corfu, Greece

³ Department of Informatics, Athens University of Economics and Business, Athens, Greece
{d.gavrilis,c.papatheodorou,p.constantopoulos,s.angelis}@dcu.gr

Abstract. This paper presents Mopseus, a Fedora-commons based digital repository that focuses on preservation. An overview of the general architecture of the system is presented along with some more in-depth details of its semantic structures. Mopseus features dynamic RDF- based relations, a service for defining metadata schemas, a built-in RDBMS synchronization and indexing mechanism, a mechanism for migration from existing repositories and a built-in workflow engine.

Keywords: Digital libraries, repository, digital preservation.

1 Introduction

Nowadays several platforms have been developed to support the creation of digital repositories. However a few of them focus on preservation and facilitate the repository administrators to implement preservation plans. The existing preservation platforms, such as CASPAR [1] and Planets [2], provide infrastructures to meet the requirements for preservation actions of large memory organizations such as national libraries and archives. A crucial issue is how much effort users are required to put in order to develop digital repositories on top of such platforms, especially when these users are small institutions with tight, small budgets [3]. Existing repository platforms, such as eSciDoc (<http://www.escidoc.org/>), offer a number of powerful services, while some, such as Blacklight (<http://www.projectblacklight.org/>), offer an easy interface. However they are complex for small – medium organizations and/or demand a number of pre-requisites to be setup.

This poster presents Mopseus, a digital library service, inspired by the conceptualization of [4] and built on top of Fedora-commons middleware that provides repository development and management services in combination with basic preservation workflows and functionalities. Mopseus is designed to facilitate medium and small institutions to develop and preserve their own repositories [5]. In comparison to the Fedora-commons platform, Mopseus provides a ready-to-use repository system, without the need of customization and the programming workload that Fedora-commons involves. In Mopseus indexing is efficiently performed using a RDMS. Additionally, a major characteristic of Mopseus is its ease of installation and development of front and back ends.

2 Architecture

Mopseus' architecture aims at flexibility, simplicity and preservation. The core of Mopseus is implemented as a set of Java services. Furthermore, an API, developed in PHP, allows for rapid development of back and front-end functionalities. The content of a Mopseus repository is stored as *digital objects*, consisting of *datastreams*, which can be text/xml, text/rdf or binary. Thus datastreams can be correlated to form digital objects that are structures of data and metadata. A digital object may be correlated to one or more *containers* using various types of relations. A container may include digital objects or other containers as well. Each Mopseus digital object is an instance of one of the following *namespaces*:

- **config:** The configuration of the repository itself is encoded by and stored as digital objects of this namespace. This makes Mopseus a self-describing repository, which means that all information regarding the setup of the repository is stored as digital object itself and thus is preserved following homogeneous and common preservation mechanisms.
- **cid:** This namespace contains digital objects that describe containers. The links between items and containers are defined using RDF.
- **iid:** This namespace contains all the digital objects that carry actual information, consisting of datastreams.
- **trm:** This namespace contains all digital objects that carry terminology information.

The Mopseus components, shown in figure 1, are:

- **Dynamic definition of XML schemas.** Mopseus provides a service for the definition of metadata schemas. The service supports the development of an XML schema, which defines the syntax of the metadata elements, their functionality (mandatory/optional elements) and presentation. This XML schema is automatically transformed in HTML forms where the user can ingest metadata and produce valid XML documents stored as datastreams. It is worth noting that multiple metadata schemas can co-exist in the same digital object, providing a different perspective than that of CASPAR Preservation Data Store [1].
- **RDBMS Synchronization.** A mechanism was developed to dynamically synchronize all the elements of the hosted XML and RDF datastreams with an external RDBMS database (currently MySQL). This process drastically improves the efficiency and flexibility of the indexing mechanism by allowing the administrator to map elements from the XML and RDF datastreams using XPath queries to the corresponding RDBMS table elements.
- **Mapping between XML schemas.** This mechanism allows the mapping between metadata schemas.
- **Workflow engine.** The workflow engine allows for easy automation of simple tasks such as *ingestion*, *revision*, etc. Each workflow is encoded as an XML document, while a graphical interface guides the user to complete the task.
- **Terminology service.** The terminology service allows for management of vocabularies, which can then be used in metadata schemas.

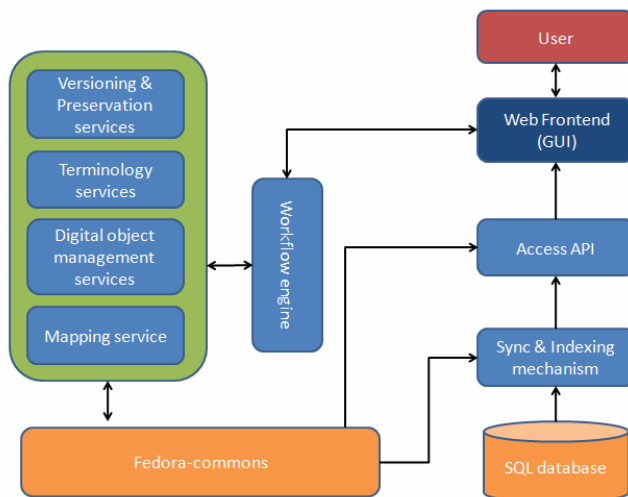


Fig. 1. Mopseus architecture

Mopseus enables the flexible definition of the semantic relationships between the objects of a repository. In particular it provides an ontology that defines:

- object groups, named *containers*; a container may contain digital objects or other containers, while a digital object may belong to more than one containers;
- relations between digital objects, datastreams and containers

Fedora commons allows using RDF-based semantic relations between digital objects. Mopseus exploits this capability to support ontologies that define the relations between the digital objects as well as between the structural parts of a digital object (i.e., its datastreams) and treat them as preservable objects themselves. This is done by defining them in RDFS and storing them in datastreams. Specifically, two datastreams residing in the config namespace have been implemented:

- RELS-EXT. Contains an ontology for describing relationship types between digital objects and containers, such as *isMemberOfCollection*, *isPartOf*, etc.
- RELS-INT. Contains an ontology for characterizing the constituent datastreams of a digital object and the relationships between them e.g. *isDocument*, *isThumbnail*.

3 Preservation Features

Mopseus is inspired by the OAIS model principles, in the sense that (a) the digital objects carry meaningful information about their binary content and relationships and (b) this representation information constitutes itself a digital object. Each digital object, as well as its relationships with other objects, is described in Mopseus by a set of datastreams, each of them being versionable. Furthermore, for every submission, a new version is created and all the previous versions are stored in Fedora's FOXML format [6]. Hence subsequent changes to XML schemas are versioned.

Moreover, Mopseus supports ingestion, access, storage, data management, administration and preservation planning OAIS functionalities, complementing the Planets workflow engine [2]. The ingestion/modification workflows are described by XML documents. Regarding preservation planning, Mopseus provides a migration process from existing repositories, facilitated through the use of a desktop tool implemented in Java. Currently it supports migration from DSpace repositories.

One of the most representative installations of Mopseus is Pandemos, the digital library of Panteion University, Athens, Greece (<http://library.panteion.gr/pandemos>). Originally, Pandemos was a DSpace repository, holding approximately 2200 digital objects, migrated to Mopseus without any loss of information and at least 5000-5500 new digital objects were ingested.

4 Conclusions and Future Work

Mopseus is a powerful and easy to configure repository management open source software, enhanced by preservation functionalities, that enables small and medium memory organizations to manage their digital holdings. Among the plans for Mopseus further development is the incorporation of relations in the ontology that denote the provenance of the digital content. Moreover, adding new workflow wizards to the workflow engine and introducing automated expression of the stored information in terms of the PREMIS metadata schema, are some of the future tasks towards the enhancement of Mopseus as a preservation tool.

References

1. Giaretta, D.: The CASPAR Approach to Digital Preservation. *The International Journal of Digital Curation* 2(1) (2007), <http://www.ijdc.net/ijdc/article/view/29/32>
2. King, R., Schmidt, R., Jackson, A., Wilson, C., Steeg, F.: The Planets Interoperability Framework - An Infrastructure for Digital Preservation Actions. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) *ECDL 2009*. LNCS, vol. 5714, pp. 425–428. Springer, Heidelberg (2009)
3. Roberts, G.: Small Libraries, Big Technology. *Computers in Libraries* 25(3), 24–26 (2005)
4. Meghini, C., Spyratos, N.: Viewing Collections as Abstractions. In: *DELOS Conference 2007: Procs of the 1st International DELOS Conference*, pp. 207–217 (2007)
5. Angelis, S., Constantopoulos, P., Gavrilis, D., Papatheodorou, C.: A Digital Library Service for the Small. In: *DigCCurr. 2009: Procs of the 2nd Digital Curation Curriculum Symposium: Digital Curation Practice, Promise and Prospects* (2009), <http://www.ils.unc.edu/digccurr2009/>
6. Fedora Object, XML (FOXML), <http://www.fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html>

Link Proximity Analysis - Clustering Websites by Examining Link Proximity

Bela Gipp^{1,2}, Adriana Taylor¹, and Jöran Beel^{1,2}

¹ UC Berkeley, Berkeley, California, USA

² Otto-von-Guericke University, Computer Science/ITI/VLBA-Lab, Magdeburg, Germany
{gipp, aitaylor, beel}@berkeley.edu

Abstract. This research-in-progress paper presents a new approach called Link Proximity Analysis (LPA) for identifying related web pages based on link analysis. In contrast to current techniques, which ignore intra-page link analysis, the one put forth here examines the relative positioning of links to each other within websites. The approach uses the fact that a clear correlation between the proximity of links to each other and the subject-relatedness of the linked websites can be observed on nearly every web page. By statistically analyzing this relationship and measuring the amount of sentences, paragraphs, etc. between two links, related websites can be automatically, identified as a first study has proven.

Keywords: Web page, Website, clustering, Network Analysis, Link Analysis, Citation Proximity Analysis.

1 Introduction

Most modern search engines offer a “find similar pages” function which returns web pages similar to a given one. Would it not be useful if an author’s knowledge of his subject matter could be used to identify similar pages?

Websites usually address a specific topic, and each section addresses a particular facet of that topic. Embedded hyperlinks operate in a parallel manner; the closer two links are to each other, the more likely it is that they have a similar theme. Since web pages necessarily reflect the course of human cognition, the adjacency of links is also an indication of the author’s conception of their relatedness, and likely of the user’s perception of their similarity, as well. The approach presented here seeks to exploit this by analyzing link proximity to identify related web pages.

2 Related Work

The main strategies for assessing the similarity of hypertext documents are text-based, user-based, and link-based analyses [9, 10]. While their synchronous operation is the end goal, the focus of this paper is on improving the last.

Link-based techniques have the advantage of circumventing text-based methods’ dependency on a document’s language, ambiguous nomenclature, synonyms and

homonyms [5]. Furthermore, the application of customary measures of likeness, sc., cosine and extended Jaccard similarity, has been straightforward [8].

Cluster analysis has been used to aid in similarity search on the web. Agglomerative and divisive hierarchical clustering algorithms; partitional clustering algorithms, like *k*-means; and density-based clustering algorithms have all been used to partition the web, in combination with link-based analysis [6].

Naturally, there are various link-based systems that determine hypertext documents' similarity to each other. Co-citation analysis and bibliographic coupling are two such means. Co-citation, as advanced by [4] and [7], occurs when two documents are cited by another document. Bibliographic coupling determines correspondence via the number of citations that two documents have in common [3].

Traditional link-based approaches like those do not take into account the internal structure of the web pages in question. The use of Citation Proximity Analysis (CPA) [1] would help in this regard. CPA adds another dimension to Co-citation by accounting for the joint appearance of citations with respect to their mutual proximity. The key presupposition is that citations have a greater probability of being related the closer they are to one another. CPA is a finer tool overall, and has the advantage of being able to determine the relatedness of subsections of documents.

We have not found CPA's underlying concept used in a link-based approach to similarity search in hypertext documents. It has previously been applied only to scientific articles, where it delivered good results [1].

3 Link Proximity Analysis

The following example is a screenshot from a Wikipedia article about the 25 largest daily newspapers in America.



Fig. 1. Example of website with links

Figure 1 illustrates that links on websites are usually the more related the closer they are listed to each other. Whereas the link to the “Audit Bureau of Circulation” is only to some extent related to the “Wall Street Journal” (see arrow *a*), the other entries listed before and after the New York Times (see arrow *b*) are closely related - all of them are newspapers.

In LPA, this fact is used to calculate the link proximity and so to identify related websites. If two links are given within one sentence they are probably addressing a

similar topic. If, on the other hand, two links are separated by a whole paragraph, then they likely address less related topics. Unfortunately, most websites do not list information in as structured a way as Wikipedia. Nevertheless, by analyzing not one source (i.e., one website), but millions, and by only considering the most frequent link combinations, outliers such as the combination of “joint operation agreement” and “Wall Street Journal” (see Figure 1) are not significant enough to be considered.

So far, we have used a simple weighting approach that only considers the amount of words, sentences, paragraphs and section headings between links. We ignored factors like different fonts, font sizes etc. If links were ordered alphabetically, the position within the list was ignored. If more than one web page linked the URI, we used the average.

Algorithm: Calculate Link Proximity

Input: Crawled websites w_i containing links l_j , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, m$

Output: Related websites

// to assign Link Proximity Index to pairs of links within a page for all web pages in w_i :

$\exists l_j \forall w_i$: // Exclude pairs of identical links

if distance d = same sentence, then LPI = 1,

if d = same paragraph, then LPI = $\frac{1}{2}$,

if d = same section, then LPI = $\frac{1}{4}$ etc.

// to calculate the average LPI for a given pair of links:

$$\forall (l_a, l_b) \in w_i,$$

$$\sum_{k=1}^x LPI(l_a, l_b)$$

$$\frac{\sum_{k=1}^x LPI(l_a, l_b)}{x}, \text{ where } x \text{ is the number of pairs of } (l_a, l_b) \in w_i$$

A first empirical study was conducted to evaluate the performance of this approach. The target websites were chosen from the most highly trafficked sites¹. To determine related pages, the structures of 500,000 sites linking the targets were analyzed. Stimuli were culled from the top 50 targets where the outlined algorithm returned a suggestion different from Google’s. In 552 cases, according to 20 volunteer test subjects, Google delivered better “related web page” results, whereas in 448 cases the described approach delivered superior results. Usually, the best recommendations are generated by hybrid approaches such as combining text, user behavior and link analysis [2]. It seems likely that this is also true for LPA.

We plan to expand this study with the involvement of interested researchers to compare the performance of LPA with existing text-, link- and user behavior-based approaches. In order to facilitate this, we released the crawler and LPA-Software as Open-Source under the General Public License.

4 Conclusion

In this paper we proposed a new approach to identify related websites based on link analysis. In contrast to the traditional approaches, it additionally analyzes the proximity of links to each other within websites. A first study showed that this approach leads

¹ According to Alexa.com

to good results despite its simplicity. However, a more comprehensive and comparative study needs to be done to evaluate the potential and fields of applications such as movie recommender systems etc.

References

- [1] Gipp, B., Beel, J.: Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In: Larsen, B., Leta, J. (eds.) Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI 2009), Rio de Janeiro, Brazil, vol. 2, pp. 571–575 (July 2009), ISSN 2175-1935, <http://www.sciplcore.org>
- [2] Gipp, B., Beel, J., Hentschel, C.: Scienstein: A Research Paper Recommender System. In: Proceedings of the International Conference on Emerging Trends in Computing (ICETiC 2009), Virudhunagar, India, pp. 309–315. Kamaraj College of Engineering and Technology India/IEEE (January 2009), <http://www.sciplcore.org>
- [3] Kessler, M.M.: Bibliographic coupling between scientific papers. *American Documentation* 14, 10–25 (1963)
- [4] Marshakova, I.V.: System of document connections based on references. *Scientific and Technical Information Serial of VINITI* 6(2), 3–8 (1973)
- [5] Fogaras, D., Rácz, B.: Scaling link-based similarity search. In: Proceedings of the 14th International Conference on World Wide Web Conference (2005)
- [6] Dutta, A.K.R., Ghosh, I., Mukhopadhyay, D.: An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure. *Journal of Convergence Information Technology* 4, 65–71 (2009)
- [7] Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science* 24, 265–269 (1973)
- [8] Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: Workshop on Artificial Intelligence for Web Search (AAAI 2000), pp. 58–64 (2000)
- [9] Klein, D., Haveliwala, T.H., Gionis, A., Indyk, P.: Evaluating strategies for similarity search on the web. In: Proceedings of the 11th International Conference on World Wide Web (2002)
- [10] Wang, Y., Kitsuregawa, M.: Evaluating contents-link coupled web page clustering for web search results. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, p. 506. ACM, New York (2002)

SliDL: A Slide Digital Library Supporting Content Reuse in Presentations*

José H. Canós, María Isabel Marante, and Manuel Llavador

Dept. of Computer Science (DSIC), Technical University of Valencia,
Camino de Vera s/n, E46022, Valencia, Spain
{jhcanos,mmarante,mllavador}@dsic.upv.es

Abstract. Presentation building applications lack good support to slide reuse. In this paper, we introduce SliDL, a digital library that facilitates slide reuse by flattening the presentation-based structure of current systems and providing slide retrieval facilities. The service-oriented architecture of SliDL enables slide sharing between different applications. We have developed clients for Microsoft PowerPoint 2007 and OpenOffice.org Impress.

Keywords: Slide reuse, presentation management, Service-Oriented Architecture.

1 Introduction and Motivation

Content reuse is one of the unsolved matters of current presentation tools. Although systems like SlideShare¹ allow sharing presentations, their presentation-centered approach makes it difficult to reuse slides belonging to other presentations during the preparation of a new one. To reuse a slide, a user must know the presentation it belongs to, open the presentation, look for the slide, and copy and paste it when found. This becomes a serious drawback when the amount of presentations is large. Only the newest version of Microsoft PowerPoint includes a feature to support slide reuse: Slide Libraries² can be used to export slides, which can be included in further presentations. This feature requires the computer to be connected to a server running Microsoft Office SharePoint Server 2007, and in this case, only slides created with PowerPoint can be reused. Consequently, a PowerPoint user cannot find nor import slides created with other programs such as OpenOffice.org Impress. Moreover, slides can only be found via browsing the entire server library, without support to neither content-based nor metadata-based searches on clients.

* SliDL has been implemented by students of the 2008-2009 edition of the Digital Libraries course of the “Master en Ingeniería del Software, Métodos Formales y Sistemas de Información” at the Universidad Politécnica de Valencia. The work of J. H. Canós and M. Llavador is partially funded by the Spanish Ministerio de Educación y Ciencia (MEC) under grants META (TIN2006-15175-C05-01) and CAPES/DGU 2008 (PHB2007-0064-PC), and Generalitat Valenciana (ACOMP07/216). M. Llavador is the holder of the MEC-FPU grant #AP2005-3356.

¹ www.slideshare.net

² <http://office.microsoft.com/en-us/sharepointserver/HA101741171033.aspx#1>

SliDL is a Slide Digital Library developed to reduce the difficulties of slide reuse. The overall goal of SliDL is to help users to build presentations by reusing previously generated slides contained in some presentation files. This means that it must provide services for storing and retrieving slides. The following features are included in the current release:

1. Platform independence: users are able to store presentations and reuse slides generated with different programs. A platform independent slide repository stores slides created with different applications.
2. Cross-platform usage: SliDL has been implemented following a service-oriented approach, so that the storage and retrieval services can be shared by every client application.
3. Slide metadata management: some slide properties are stored to allow metadata-based retrieval.
4. Content-based slide retrieval: the textual content of slides is indexed to allow keyword-based slide search.
5. Slide collection browsing: SliDL users can browse the thumbnail collection to find slides. Browsing can be made over all the slide collection or over the results of a previous content and/or metadata based search.

We describe the architecture of SliDL and the services provided in its current release. We also outline some of the distinguished features of the forthcoming release.

2 The Service-Oriented Architecture of SliDL

Figure 1 shows the layered architecture of SliDL. At the Storage Layer, the SliDL Repository stores the slide collection. The repository is updated and queried through Indexing and Searching components of the SliDL API at the Service Layer. Presentation software at the client layer will use the services of the SliDL API to send and retrieve slides information to and from the repositories. We give a more detailed description of each layer in the following subsections.

2.1 Client Layer

The Client Layer is where the presentation software is placed in SliDL. In order to be able to interact with the SliDL services, clients for the different programs must be developed. A client is a small program able to handle a presentation using the standard application's object model (i.e., a plug-in for the specific application). A SliDL client has two main components (see Fig 1). On one hand, the *Slide Extractor* processes a presentation and extracts its slides; then, it uses the services of the SliDL API to send the slides, along with their associated metadata, to the repository.

On the other hand, slide retrieval facilities are provided by the *Slide Seeker*. The module offers several ways to retrieve slides:

- Metadata-based search: typical searches based on author, title, date, etc.
- Content-based text retrieval: the user can type one or more words, and the slide seeker returns a list of the slides containing such words in some of their textual shapes.

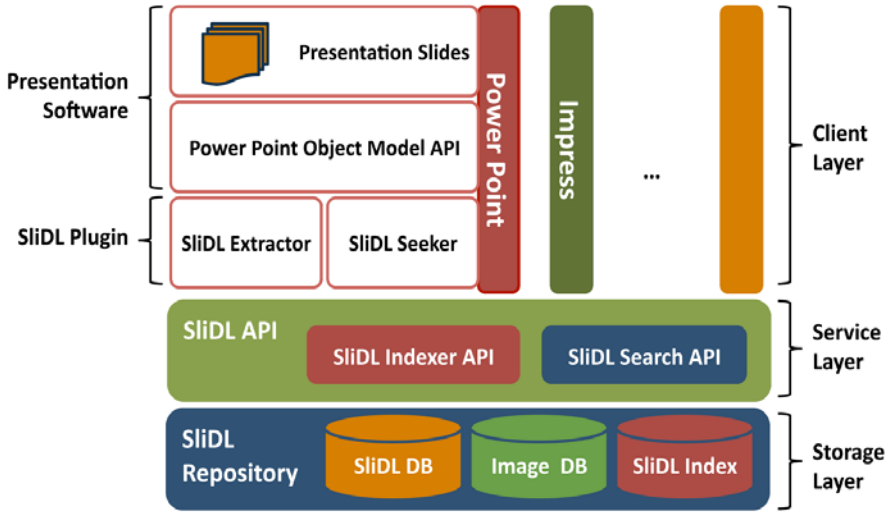


Fig. 1. Architecture of SliDL with details about SliDL clients

- Thumbnail-based slide browsing: the user may select the requested slide clicking on the corresponding slide.
- Mixed strategies: a user can first make a content-based query, and then browse the thumbnails of the returned slides to precisely locate the slide sought.

2.2 Service Layer

The service layer offers services for storing and retrieving slides from the SliDL repository, which is accessed via a service-oriented API (<http://mborges.dsic.upv.es:8001/SLIWS/service.asmx>). To provide interoperability, we have defined a common slide exchange format. The slides sent/obtained by clients to/from the repository are encoded in XML.

2.3 Storage Layer

The SliDL Repository holds the collection of slides; for each slide, the following information is held:

- Descriptive and structural metadata coming from both the slide and the presentation it belongs to (e.g. author, title of presentation, title of slide); slides metadata are stored in the SliDL DB component.
- Text content: all the text contained in textual shapes, including headers and footers.
- Thumbnail: a small image of the slide. The purpose of thumbnails is twofold. On one hand, they can be used to browse through the slide collection; on the other hand, they can be used in the future to use some image-based slide retrieval facilities.

The process of storing a presentation in the Repository is as follows: the SliDL Extractor of a client processes a presentation every time it is saved after some change³. For each slide in the presentation, index terms are extracted from the content of text shapes, and put into the SliDL Index; in addition, a thumbnail of the slide is generated and added to the Image DB; finally, metadata is extracted from the slide and the container presentation and inserted into the SliDL DB.

3 Current Status and Further Work

The current version of SliDL is for personal use, that is, without support to multiple users. Its goal is to help users to organize their presentations. We have developed clients for both Microsoft PowerPoint 2007 and OpenOffice.org Impress.

We are moving to a new version where the following features will be available:

- Multi-user facilities: SliDL will allow users to define a personal space within the digital library and decide whether the content of such space must be considered private or accessible to other users of the system.
- Slide tagging and rating to enhance the retrieval facilities.
- Tracking and reviewing changes to slides on the server to keep user's presentations updated.
- Duplicate slide detection to find slides included in different presentations.
- Metadata-based recommender systems: exploiting relationships such as "belong to the same presentation" may enrich the search results; also, as a mid-term goal, it would be desirable to build systems able to suggest a preliminary presentation from a set of keywords using the slides contained in SliDL.

³ It is possible to explore folders and process all presentation files found as an iteration of the described presentation indexing process.

Metadata Impact on Research Paper Similarity

Germán Hurtado Martín^{1,2}, Steven Schockaert^{2,*},
Chris Cornelis^{2,*}, and Helga Naessens¹

¹ Dept. of Industrial Engineering, University College Ghent, Belgium

² Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium

Abstract. While collaborative filtering and citation analysis have been well studied for research paper recommender systems, content-based approaches typically restrict themselves to straightforward application of the vector space model. However, various types of metadata containing potentially useful information are usually available as well. Our work explores several methods to exploit this information in combination with different similarity measures.

1 Introduction

Given the proliferation of published research results, recommending scientific papers to researchers may provide a useful complement to traditional literature search [14, 5]. Various approaches may be taken to automate this task, including collaborative filtering (e.g. based on CiteULike.org or Bibsonomy.org), citation analysis (e.g. PageRank, HITS, etc.) and content-based (CB) approaches. In this paper, we focus on the latter type of systems.

Typically, CB approaches use cosine similarity applied to tf-idf vector representations of the abstracts for comparing research papers. However, various kinds of metadata are usually associated with papers, including keywords, scientific classification, journal of publication, etc.; to our knowledge, their impact on identifying related papers has not been investigated previously. In this paper, therefore, we perform an exploratory study of various methods which use such metadata directly or indirectly. Apart from assessing the relative worth of the various methods, our findings also serve to set out a baseline for future work on CB paper recommendation strategies.

2 Methodology

Test collection. To build a test collection for evaluating similarity measures, we crawled a portion of the ACM library [1], consisting of all articles from 23 journals in the Artificial Intelligence domain. In addition to abstract, title, authors and journal, we also extracted the entries from the ACM classification system that

* Postdoctoral fellows of the Research Foundation – Flanders.

¹ <http://portal.acm.org>

were assigned to the paper, its general terms (taken from a fixed thesaurus), keywords (freely chosen by the authors) and cited papers. A description of 34658 papers was thus retrieved. Our experiments are restricted, however, to the 9594 papers for which none of the extracted fields is empty.

Similarity measures. The most straightforward way to measure the similarity between two papers is by comparing their abstracts in the vector space model (method *abstract* in Table [1](#)); each paper is represented as a vector, in which each component corresponds to a term occurring in the collection. The value of that term is calculated using the standard tf-idf approach, after removing stop words. The vectors \mathbf{p} and \mathbf{q} corresponding to different papers can then be compared using standard similarity measures such as the cosine (*cos* in Table [1](#)), generalized Jaccard (*g.jacc*), extended Jaccard (*e.jacc*), and Dice (*dice*) similarity, defined respectively by

$$\begin{aligned} sim_c(\mathbf{p}, \mathbf{q}) &= \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\| \cdot \|\mathbf{q}\|} & sim_{gj}(\mathbf{p}, \mathbf{q}) &= \frac{\sum_k m_k}{\sum_k M_k} \\ sim_{ej}(\mathbf{p}, \mathbf{q}) &= \frac{\sum_k m_k}{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - (\mathbf{p} \cdot \mathbf{q})} & sim_d(\mathbf{p}, \mathbf{q}) &= \frac{2(\mathbf{p} \cdot \mathbf{q})}{\|\mathbf{p}\|^2 \cdot \|\mathbf{q}\|^2} \end{aligned}$$

where $\mathbf{p} \cdot \mathbf{q}$ denotes the scalar product, $\|\cdot\|$ the Euclidean norm, $m_k = \min(p_k, q_k)$, and $M_k = \max(p_k, q_k)$. Alternatively, papers can be represented as vectors whose components refer to the general terms (method *g.terms*), to the keywords (method *keywords*), or to the classes of the ACM classification that have been assigned to it (method *class*). The weights are calculated analogously as in the tf-idf model. To cope with the tree structure of the ACM classification, in the *class* method, we do not only add a component for the classes at the lowest level, but also for each of their ancestors; tf-idf weighting then ensures that more emphasis is put on the lower level classes.

In the previous methods, metadata such as classes or keywords is used directly, in such a way that the most important information, the abstract, is completely ignored. We therefore follow an alternative scheme, which we refer to as explicit semantic analysis (ESA) since it is analogous to the approach from [2](#). Let \mathbf{p} be the vector representation obtained by method *abstract*. We now define a new vector representation \mathbf{p}_E of this paper, with one component for every keyword k appearing in the collection. To define the weights of \mathbf{p}_E 's components, a new collection $\mathcal{C}_E = \{\mathbf{q}_k | k \text{ is keyword}\}$ is first created, where \mathbf{q}_k is a vector representation of the concatenation of the abstracts of all papers to which keyword k was assigned. The weights in vector \mathbf{q}_k are the tf-idf scores calculated w.r.t. the new collection \mathcal{C}_E . The weight w_k in \mathbf{p}_E corresponding to keyword k is then defined by

$$w_k = \mathbf{p} \cdot \mathbf{q}_k$$

This method is called *ESA-kw*. Similar methods are considered in which vector components refer to authors (*ESA-aut*) or to classes (*ESA-cl*). For efficiency,

only authors are considered that appear in at least 4 papers in the *ESA-aut* method, and only keywords that appear in at least 6 papers in the *ESA-kw* method.

Evaluation metrics. The ground truth for our experiments is derived from citations. In particular, we consider two papers as similar if either of them has cited the other, and not similar otherwise. To evaluate the performance of the methods, each paper \mathbf{p} is compared against 13 others that were published in the same journal, 3 of which are actually considered similar. Similarity measures can then be used to rank the 13 papers, such that ideally the papers similar to \mathbf{p} appear at the top of the ranking. In principle, we thus obtain one ranking per paper in the collection. However, since some papers are not sufficiently cited by papers that are also in the collection, only 3758 rankings were actually obtained. Their rankings can then be evaluated using standard information retrieval metrics; we use mean average precision (MAP) and mean reciprocal rank (MRR).

3 Results and Discussion

Table 1 summarizes the results of the experiment. A first important conclusion is that a content-based approach to finding related papers appears to be reasonable, as witnessed by the relatively high MAP and MRR scores of the best performing configurations. Another obvious conclusion is that all the other methods are worse or comparable to the traditional approach, *abstract*, although surprisingly the generalized Jaccard performs significantly better than the popular cosine method (paired t-test, $p < 0.001$). On the other hand, except for *g.terms* all of the methods perform substantially better than random (MAP 0.367, MRR 0.453). As could be expected, general terms are not sufficiently focused to help finding related papers. The keywords and ACM classification do seem to be useful, although alone they cannot beat *abstract*. Intuitively, keywords may be too specific, and the ACM classes too general to derive more accurate similarity information. It therefore seems promising to investigate methods that combine ACM class information with available keywords. Future work will also focus

Table 1. Experimental results

	MAP				MRR			
	<i>cos</i>	<i>dice</i>	<i>e.jacc</i>	<i>g.jacc</i>	<i>cos</i>	<i>dice</i>	<i>e.jacc</i>	<i>g.jacc</i>
abstract	0.581	0.581	0.581	0.594	0.724	0.723	0.723	0.741
g.terms	0.367	0.367	0.368	0.367	0.443	0.443	0.444	0.442
keywords	0.472	0.469	0.470	0.475	0.634	0.631	0.629	0.634
class	0.432	0.430	0.429	0.420	0.545	0.543	0.538	0.528
ESA-aut	0.505	0.505	0.505	0.518	0.643	0.643	0.643	0.674
ESA-cl	0.535	0.535	0.535	0.527	0.667	0.667	0.667	0.673
ESA-kw	0.597	0.597	0.597	0.553	0.748	0.748	0.748	0.704

on improving the *keywords* and *class* methods by taking dependencies among keywords/classes into account (e.g. based on fuzzy rough sets, as proposed in [3]).

The ESA methods, in general, seem to outperform their “classical counterparts”. However, these methods are computationally considerably more demanding, and while they make use of the abstract information, they do not succeed in improving *abstract* substantially. These results therefore suggest that the ideas behind ESA are not particularly suitable in this context. However, initial results indicate that the relative performance of the ESA methods strongly depends on the size of the test collection. In future work we will investigate the influence of the size of the test collection in more detail, as well as the role of the specific evaluation task. For instance, ground truth can be obtained using other methods than citation. An interesting idea is to derive it from user profiles from CiteULike as in [1].

As several types of metadata clearly show potential, it seems promising to consider methods for automatically learning to rank papers, based on a combination of abstract information and other features.

References

1. Bogers, T., Van den Bosch, A.: Recommending scientific articles using CiteULike. In: Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys 2008), pp. 287–290 (2008)
2. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606–1611 (2007)
3. Hurtado Martín, G., Cornelis, C., Naessens, H.: Personalizing information retrieval in CRISs with Fuzzy Sets and Rough Sets. In: Proceedings of the 9th International Conference on Current Research Information Systems (CRIS 2008), pp. 51–59 (2008)
4. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the recommending of citations for research papers. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW 2002), pp. 116–125 (2002)
5. Proceedings of ECML PKDD (The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases) Discovery Challenge 2009, Bled, Slovenia (September 7, 2009)

Exploring the Influence of Tagging Motivation on Tagging Behavior

Roman Kern¹, Christian Korner², and Markus Strohmaier^{1,2}

¹ Know-Center, Graz

² Graz University of Technology

rkern@know-center.at, {christian.koerner,markus.strohmaier}@tugraz.at

Abstract. The reasons why users tag have remained mostly elusive to quantitative investigations. In this paper, we distinguish between two types of motivation for tagging: While *categorizers* use tags mainly for categorizing resources for later browsing, *describers* use tags mainly for describing resources for later retrieval. To characterize users with regard to these different motivations, we introduce statistical measures and apply them to 7 different real-world tagging datasets. We show that while most taggers use tags for both categorizing and describing resources, different tagging systems lend themselves to different motivations for tagging. Additionally we show that the distinction between describers and categorizers can improve the performance of tag recommendation.

1 Introduction

Tags in social tagging systems are used for a variety of purposes [1]. In this paper, we study the distinction between two different tagging behaviors. The first type of tagging is similar to assign resources to a predefined classification scheme. Users motivated by this behavior use tags out of a controlled and closed vocabulary. These users, named *categorizers*, tag because they want to construct and maintain a navigational aid to resources for later browsing. On the other hand, users who are motivated by description view tagging as a means to accurately and precisely describe resources. Tags produced by this user group resemble keywords that are useful for later searching [2]. This distinction can be exploited for example to improve the performance of tag recommender systems and information retrieval applications. Figure 1 contrasts a tag cloud of a typical categorizer with a tag cloud of a typical describer.

2 Development of Measures

To characterize the extent to which users categorize or describe resources, we present statistical and information-theoretic measures that are independent of the meaning of tags, the language of tags, or the resources being tagged.

Characterizing Categorizers: The activity of tagging can also be viewed as an encoding process, where tags encode information about resources. If this would

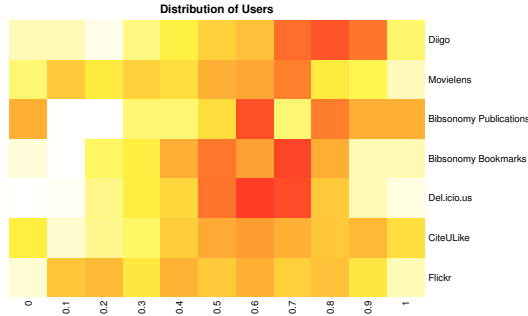


Fig. 2. Distribution of users of the real-world datasets according to the $M_{combined}$ measure. The intensity of the color encodes the relative number of users within a bin. The bins on the left side represent categorizer, while the rightmost bins represent users that display a behavior typical for describers. For example the **Flickr** dataset contains users evenly distributed between the two extremes, whereas the majority of users in the **Diigo** dataset are identified as describers.

study whether the various tagging systems differ in regard to the two user types. Figure 2 demonstrates that the distribution of describers and categorizers vary between the individual social tagging systems. For example the **Diigo** dataset contains many users that are identified as describers. One possible reason for this might be the fact, that the **Diigo** platform not only offers the possibility to tag resources, but also to create so called bookmark lists, which is better suited to categorize resources.

Tag Recommender. Finally we implemented two simple tag recommender systems to test whether the distinction between the two user groups could improve the performance of tag recommendation. The first recommender draws tags from the personal tagging history of a user and is labeled as *personomy-based recommender*. The *folksonomy-based recommender* suggests the most frequent tags as used by other describer users. Users were split into a describer and categorizer group according to the $M_{combined}$ measure. The baseline was produced by randomly assigning users to one of the two groups. For the evaluation we used the **Del.icio.us** dataset, as the folksonomy-based recommender requires resources tagged by multiple users. Figure 3 depicts the performance of the tag recommenders for different splits of the userbase (from 10% categorizers and 90%

Table 1. Overview of the size and characteristics of the crawled real-world datasets

Dataset	$ U $	$ T $	$ R $	$ R_u _{min}$	$ T / R $
Delicious	896	184,746	1,089,653	1,000	0.1695
Flickr Tags	456	216,936	965,419	1,000	0.2247
Bibsonomy Bookmarks	84	29,176	93,309	500	0.3127
Bibsonomy Publications	26	11006	23696	500	0.4645
CiteULike	581	148,396	545,535	500	0.2720
Diigo Tags	135	68,428	161,475	500	0.4238
MovieLens	99	9,983	7,078	500	1.4104

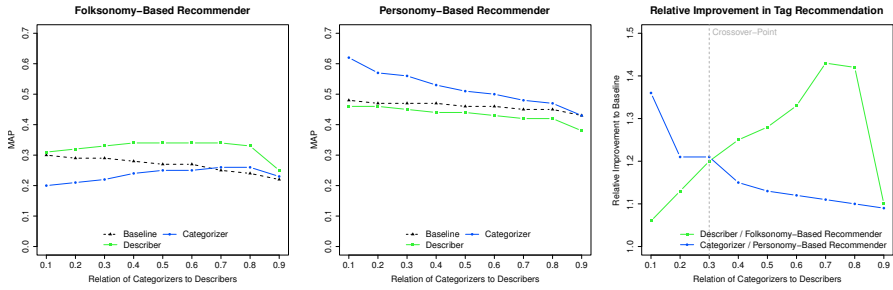


Fig. 3. Suggesting tags also used by other users appears to be a good strategy for describers (left). Categorizer prefer to reuse tags from their personal tagging history (middle). The relative improvements indicates that for about 30% of all users in our Del.icio.us dataset the personomy-based recommender is the better choice (right).

describers up to a 90%:10% split). One can see that using the personal tagging history is helpful for categorizers, while describers appear to tags similar to other users (describers) in the folksonomy. Especially of interest is the point where the relative improvement of the two recommenders intersect each other (right chart in figure 3). When developing a production tag recommender, this would be the point to switch from personomy-based tag recommendation for categorizers to a folksonomy-based recommender for describers.

4 Conclusion

We showed that different tagging systems lend themselves to different motivations for tagging. Our results reveal that even within tagging systems, tags are adopted in different ways. One of the major implications of our work is that tagging motivation exhibits significant variety, which could play an important part in a range of problems including tag recommendation and information retrieval. In previous work [4], we have demonstrated that the motivation behind tagging influences the performance of semantic acquisition algorithms in folksonomies. Improving existing state-of-the-art tag recommenders by incorporating the tagging motivation is one of the main goals of our future work.

Acknowledgements. The research presented in this work is in part funded by the Know-Center and the FWF Austrian Science Fund Grant P20269. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of VIT, Austrian Ministry of WA and by the State of Styria.

References

1. Heckner, M., Heilemann, M., Wolff, C.: Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In: Int'l AAAI Conference on Weblogs and Social Media (ICWSM), San Jose, CA, USA (2009)

2. Strohmaier, M., Körner, C., Kern, R.: Why do users tag? detecting users motivation for tagging in social tagging systems. In: International AAAI Conference on Weblogs and Social Media (ICWSM 2010), Washington, DC, USA, May 23-26 (2010)
3. Körner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: 21st ACM SIG-WEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada. ACM Press, New York (June 2010)
4. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop thinking, start tagging: Tag semantics arise from collaborative verbosity. In: Proceedings of the 19th Int'l Conf. on World Wide Web - WWW 2010 (2010)

A Teaching Tool for Parasitology: Enhancing Learning with Annotation and Image Retrieval

Nádia P. Kozevitch¹, Ricardo da Silva Torres¹, Felipe Andrade¹,
Uma Murthy², Edward Fox², and Eric Hallerman³

¹ Institute of Computing, University of Campinas, Campinas, SP, Brazil

² Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

³ Department of Fisheries and Wildlife Sciences, Virginia Tech,
Blacksburg, VA 24061, USA
{nadiapk,rtorres}@ic.unicamp.br,sansa.felipe@gmail.com,
{umurthy,fox}vt.edu, ehallerman@vt.edu

Abstract. Parasitology is a basic course in life sciences curricula, but up to now it has few computer-assisted teaching tools. We present SuperIDR, a tool which supports annotation and search (based on a textual and a visual description) in the biodiversity domain. In addition, it provides a feature to aid comparison of morphological characteristics among different species. Preliminary results with two experiments show that students found the tool to be very useful, contributing to an alternative learning approach.

1 Introduction

Parasites are responsible for human and animal diseases causing suffering and economic loss. Parasitology is a basic course in the life sciences curricula, but up to now it has few computer-assisted teaching systems.

Parasite identification and comparison is an important step in treating and combating parasitic diseases. There is no centralized application available which allows digital annotation, comparison, or identification of species. Traditionally, identification and comparison rely on using the published literature, documents, microscope analysis, and identification keys. Students use and search through different sources, trying to learn and memorize diagnostic specimen features.

Online sources are rare, and students use alternative learning resources such as personal notes and annotated images. The Laboratory of Identification of Parasites of Public Health Concern site [\[1\]](#) is an example, with a small image library about parasites, presenting their details, characteristics, endemic areas, and life cycles.

Using basic services from digital libraries (annotation and image retrieval), we present an alternative approach to teach, compare, and learn concepts about parasites. For this application, we adapted SuperIDR [\[2\]](#), a digital library system previously used in the Ichthyology domain. We evaluated the application

¹ <http://www.dpd.cdc.gov/DPDX/default.htm>

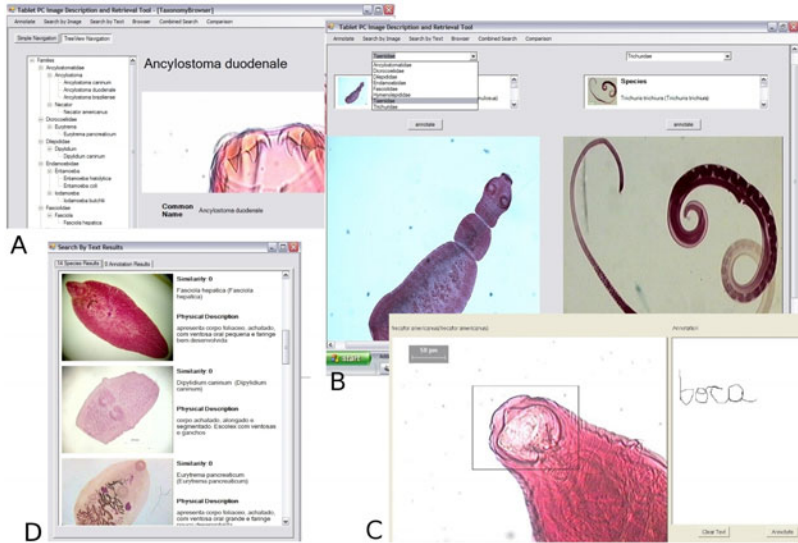


Fig. 1. SuperIDR Tool: A- Treeview navigation, B- Comparison, C- Annotation and D - Text search

with undergraduate students from the Zooparasitology class at the University of Campinas, Brazil.

2 SuperIDR

SuperIDR [\[1\]](#) combines key digital library services: text- and content-based image description and retrieval. The objective is to enhance important educational activities that involve working with specific contextualized information and sharing information among teachers and students through the use of pen-based computing, annotation, and content-based image retrieval (CBIR). Users can create marks on images, write annotations and associate them with marks and images, link text annotations with image marks, and browse and search marks and associated data (using text- and content-based retrieval mechanisms), as shown in Figure [1](#). The application also has a comparison tab, on which it is possible to load two different parasite images for a visual comparison and annotation.

Annotation is supported with the use of the tablet PC, where concepts can be related to parts of images. Each annotation is also associated with the user who created it. CBIR is supported by a content-based image search component, which supports the use of different types of vector-based image descriptors (metric and non-metric; color, texture, and shape descriptors; with different data structures to represent feature vectors).

The database included data about parasite species, listing characteristics such as species name, family, genus, similar species, mean body size, habitat, and reproductive habitat.

Table 1. Evaluation Summary

Information	Experiment 1	Experiment 2
Number of Students	13	17
Number of Species	17	25
Number of Images	76	49
Tasks	Compare images from <i>Nematoda</i> and <i>Cestoda</i> classes	Compare images from <i>Taenidae</i> class
Computer Skill	69%	92%

The main interface was developed using the .Net C# language. The content-based image search component is a Dynamic-Link Library (DLL) written in C#, resulting in a technologically flexible and portable application [2].

3 Classroom-Based Evaluation

The experiment sessions were divided into three phases: (i) the students were introduced to SuperIDR; (ii) all functionalities were tested using a manual; and (iii) students made annotations for two different species. The objective was to verify if the application helps to reinforce morphological concepts and to compare characteristics among different species.

There were two experiment sessions with students of Zooparasitology, offered by the Biology Department at the University of Campinas. The first experiment session was in 2008 [3]; the second was in 2009. Table 1 summarizes the first and second experiment regarding the number of students, images, species, tasks, and computer skills from the students.

We presented two questionnaires to the students. The first questionnaire assessed familiarity with computers, the English language, and tablet PCs. The second questionnaire sought feedback on the usefulness of the application.

Thirteen students completed and returned the initial and final questionnaires for the first experiment session. 69% of the students considered themselves as having medium or high computer skill. Students had already used annotations, but only with paper, scientific publications, and printed figures. No application known by the students allowed digital annotation, comparison, or identification. Literature, papers, microscope analysis, and bibliography identification keys were the most-used sources for identifying and comparing species. 92% of the students considered SuperIDR very useful for species identification, contributing to annotation and comparison.

The second experiment was conducted in 2009. The students were introduced to SuperIDR and used the comparison tab to insert annotations for different species. 92% of the students considered themselves as having medium or high computer skill. 100% of the students considered SuperIDR very useful for species identification, as a dynamic and interactive application. The possible explanation for a better result in the second experiment is that the application was more adapted to the parasite domain, biology professors and trainers already know

how to conduct better the experiment, and the activities were more focused on the comparison task.

Several meetings were necessary to understand some of the main concepts of the parasitology domain and to identify how the application could help the students and professors. For example, it is necessary to store different information (like shape, characteristics and figures) for each life-cycle phase of the parasite.

Students felt that the application could be improved in the following ways: (i) SuperIDR had few species; (ii) the application had few images for each species; (iii) the students were not familiar with the tablet PC, and (iv) the tablet PC recognized only English words. A possible explanation is that the number of classes were too few and the majority of the students had never worked with a tablet PC before.

Still, most of the students stated that SuperIDR was very useful as an alternative approach for learning, and would like to use the application again. Students' feedback mentioned annotation, taxonomic navigation, and comparison tabs as preferred features. Students' comments stated that the application was a very good approach for species identification and comparison.

4 Discussion and Future Work

We presented SuperIDR as an alternative approach to use digital library services: to teach and compare species in the Parasitology domain. Students performed well with SuperIDR and it was generally well received, supporting image retrieval, annotations and significant numbers of details. Future work includes adapting SuperIDR to other disciplines, import/export capabilities, geographic annotation, and linking marks with other multimedia information.

Acknowledgments

We thank CAPES, FAPESP e CNPq - BioCORE Project, HP Technology for Teaching, and Microsoft Research.

References

1. Murthy, U., Fox, E.A., Chen, Y., Hallerman, E., Torres, R.S., Ramos, E.J., Falcão, T.R.C.: Superimposed image description and retrieval for fish species identification. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) *ECDL 2009*. LNCS, vol. 5714, pp. 285–296. Springer, Heidelberg (2009)
2. Kozievitch, N.P., Falcão, T., Torres, R.S.: A .Net Implementation of a Content-Based Image Search Component. In: *Sessão de Demos, XXIII Simpósio Brasileiro de Banco de Dados*, Campinas, Brasil (2008)
3. Kozievitch, N.P., Torres, R.S., Falcão, T., Ramos, E., Andrade, F., Allegretti, S.M., Ueta, M.T., Madi, R.R., Murthy, U., Fox, E.A., Chen, Y., Hallerman, E.: Evaluation of a Tablet PC image annotation and retrieval tool in the parasitology domain. Technical Report IC-09-23, Institute of Computing, University of Campinas, Brazil (July 2009)

Framework for Logging and Exploiting the Information Retrieval Dialog

Paul Landwich, Claus-Peter Klas, and Matthias Hemmje

FernUniversität in Hagen

Institute for Multimedia and Internet Applications

{paul.landwich,claus-peter.klas,matthias.hemmje}@fernuni-hagen.de

Abstract. In this paper we present first results for a new approach of an innovative user interface for digital library and information retrieval systems. The leading thought bases on the fact that only the dialog between user and system can establish a necessary information context in order to satisfy an information need. We introduce a framework for information retrieval systems to handle activities and sets elaborated during a search process and a prototype tool integrated in an existing interface framework. Finally a description of a user study and expert interviews and their evaluation results conducted on the basis of the tool is given.

1 Introduction

When we speak about information retrieval (IR), we must distinguish between two needs of information seeking. One need is the quick retrieve of a few especial, perhaps known, information objects (known-item-search). The other need is to collect information objects within a search task to solve a problem by reducing an information deficit. For this case the cognitive information overload of users rises with the complexity of the search task. These tasks could last days or weeks with many different queries and many result sets with hundreds of information objects. The complexity of search tasks were exposed by many papers (e.g. [1], [2] and [3]). For this, the IR and digital library (DL) community need to reduce this overload with innovative and new user interfaces and services.

In the following sections we briefly outline our idea to support the user with a tool to visualize and manage the whole search task in an easy way. We will show that our prototype corroborates our hypothesis: The user feels supported in the seeking process by reducing the cognitive overload. After the outline we introduce a framework for information retrieval and digital libraries systems. A prototype tool which bases on the new framework will be introduced. Finally we describe a first pre-testing experiment and we then present and discuss its results and draw some conclusions based on these results.

2 Framework for Information Retrieval Systems

Landwich [4] outlined in his conceptual model, that the information retrieval dialog is a cycle driven by activities. These activities and their results fill our

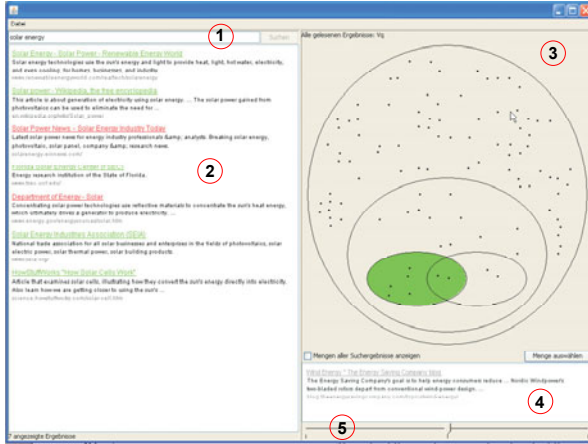


Fig. 2. Selected Set of Information Objects

it is split in five sections. Figure 2 shows a screen shot of the prototype and the numbered sections. In section 1 the user finds an input box to formulate the query. The section beneath (2) is the area to visualize the result list. Here, all elements of the displayed set are shown as a list and sorted by relevance. Clicking on the title, the corresponding website will open in the default browser. On the right upper side (3) the sets are visualized. The circles of the venn-diagramm contain elements of the different sets. Each result element is displayed as a single point. With a mouse click each set can be activated. Colours differentiate the and highlight the current set focus. With the activation of a specific set the result list on the right side will adopt to the selected set. It is also possible to choose and activate these sets and all meaningful relations through the usual context menu. When the user moves with the mouse over an element of a set, in section 4 information about this object are listed. On the bottom of the right side (5) the user finds a slider to move through the history of queries. Moving the slider to the outer right side will display the sum of all queries.

4 Evaluation

A between-subjects experimental design has been setup. With this first design we evaluated the subjective acceptance of the offered tool. The first group performed searches using the prototype tool with the underlying *Yahoo* engine. The second group performed searches using the standard interface of *Yahoo*. A questionnaire with a 5-point scale gathered information and values for this between-subjects experimental design. 15 of 21 variables show a significant difference between both groups. The results of our experiment demonstrate that the prototype tool that we designed for search processes appears to have significant advantages over a standard system like *Yahoo* which is designed to support quick searching only.

The highest difference (3,125) is measured for the question “The tool gives me control over my results”. This variable describes the main benefit of the prototype system. The user himself feels in the position to overview and to handle the search process with the elaborated sets of information objects and activities.

Three librarians and two natural scientists served as test persons. The experts could use all functions of the prototype tool and had the same introduction as all other test persons. Generally, all experts gave a positive feedback. The underlying idea to visualize and to give control over the elaborated sets of information objects was well accepted.

All results show a significant benefit in handling the growing information dialog context of a seeking process with the prototype tool and corroborate in a strong way our hypothesis.

5 Conclusions

In this paper we presented first results for a new approach of an innovative user interface for information retrieval systems. The results of the experiment demonstrated that the user feels significantly supported by the prototype system in values of support and usability. With this experiment we feel encouraged to deepen our research. In a next step we will setup an objective experimental design to measure values for the reduction of the cognitive overload. Furthermore we will refine the prototype under the aspect of the results and the made suggestions for improvement. Implementations to realize other aspects (see section 2) of our framework are planned.

References

1. Pharo, N.: A new model of information behaviour based on the search situation transition schema. *Inf. Res.* 10(1) (2004)
2. Rose, D.E.: Reconciling information-seeking behavior with search user interfaces for the web. *J. Am. Soc. Inf. Sci. Technol.* 57(6), 797–799 (2006)
3. Xu, Y.: The dynamics of interactive information retrieval behavior, part i: An activity theory perspective. *J. Am. Soc. Inf. Sci. Technol.* 58(7), 958–970 (2007)
4. Landwich, P.: Model for digital library user interfaces supporting visual information dialog services. In: *Papers from the 2008 ECDL Doctoral Consortium*, vol. 5 (2008), <http://www.ieee-tcdl.org/bulletin.html>
5. Landwich, P., Vogel, T., Klas, C.P., Hemmje, M.: Supporting patent retrieval in the context of innovation-processes by means of information dialogue. *World Patent Information* 31, 269–360 (2009), doi:10.1016/j.wpi.2009.04.004.
6. Klas, C.P., Fuhr, N., Schaefer, A.: Evaluating strategic support for information access in the DAFFODIL system. In: Heery, R., Lyon, L. (eds.) *ECDL 2004*. LNCS, vol. 3232, pp. 476–487. Springer, Heidelberg (2004)

Defining the Dynamicity and Diversity of Text Collections

Ilya Markov and Fabio Crestani

University of Lugano, Faculty of Informatics
Via G. Buffi 13, 6900, Lugano, Switzerland
{ilya.markov, fabio.crestani}@usi.ch

Abstract. In Information Retrieval collections are often considered to be relatively dynamic or diverse, but no general definition has been given for these notions and no actual measure has been proposed to quantify them. We give intuitive definitions of the dynamicity and diversity properties of text collections and present measures for calculating them based on the notion of novelty. Experimental results show that the proposed measures are consistent with the definitions and can distinguish collections effectively according to their dynamicity and diversity properties.

1 Introduction

Each time some Information Retrieval technique is extensively tested, researchers try to consider several experimental collections that are different one from the other according to a number of properties. For example, Federated Search testbeds are explicitly said to be homogeneous or heterogeneous [2], special smoothing techniques are to be applied for homogeneous collections [4], query expansion may be performed based on the terms from recent documents in relatively dynamic collections. Therefore it is important to define what the dynamicity¹ and diversity of text collections are and to be able to measure these properties quantitatively.

We consider text collections that evolve over time. This assumption comes from the real-world environment, where the documents are added to a collection or updated as time goes by (we do not consider document removal since it happens rarely nowadays). Our intuition behind measuring the *dynamicity property* of text collections is as follows. Given an evolving collection, if a newly added document is novel comparing to the documents added in the nearest past, then the collection can be considered to be relatively dynamic as opposed to a collection for which new documents are redundant comparing to the documents seen in the nearest past.

In order to measure the *diversity property* one has to compare all the documents in a collection pairwise. Based on this, one collection can be said to be more diverse than the other if, in general, its documents are more different between each other, then the documents of the other collection. Only some subset of documents in a collection can be chosen for pairwise comparison to reduce the computational cost.

¹ In the context of this paper dynamicity means the property of being dynamic.

2 Related Work

In the past years a number of research areas have been concerned with improving their results by incorporating diversity: text document retrieval [3], recommender systems [7], image retrieval [5]. As opposed to these works, our intent is to find standalone measures for estimating the diversity and dynamicity properties of text collections in general. As far as we know nobody tried to quantitatively estimate the dynamicity property of text collections.

Our intuition behind measuring the dynamicity and diversity properties of text collections is based on the notion of *document novelty*. First, the task of identifying novel and redundant documents was addressed by Zhang et al. [6]. A number of measures were proposed based on three different types of evidence: new word counts, cosine distance and distribution similarity (Kullback-Leibler divergence). It was shown that cosine distance performs the best in the task of novelty and redundancy detection for AP and WSJ TREC datasets.

Allan et al. [1] applied the measures discussed in [6] to the task of novelty detection at a sentence level. It was shown that on a sentence level word counting measures perform the best. This is because sentences with overlapping vocabulary but different word distributions are considered novel from a word distribution point of view, but redundant from a new word counting perspective.

In this work we do not consider the task of document novelty and redundancy detection itself, but apply the measures discussed in [6] and [1] to the task of calculating the dynamicity and diversity measures of text collections.

3 Measures

Each *novelty measure* receives a document d and a set of documents DS ($d \notin DS$) as an input and returns the novelty score of a given document against a given set of documents $NS(d, DS)$. The novelty measures we use include:

Average New Word Ratio: $AvgNWR(d, DS) = \frac{\sum_{d_i \in DS} |W_d \cap \overline{W_{d_i}}|}{|W_d| \cdot |DS|}$. Here W_d is a set of words in a document d .

Cosine Distance: $CosDist(d, DS) = \frac{\sum_{d_i \in DS} \cos(d, d_i)}{|DS|}$. Here V is a combined vocabulary of a document d and a set of documents DS . A document d is represented as a vector $d = (w_{1,d}, w_{2,d}, \dots, w_{n,d})^T$ and $\cos(d_1, d_2) = \frac{\sum_{t \in V} w_{t,d_1} w_{t,d_2}}{\|d_1\| \cdot \|d_2\|}$.

Average Language Model KL Divergence: $AvgLM(d, DS) = \frac{\sum_{d_i \in DS} KL(\Theta_d, \Theta_{d_i})}{|DS|}$.

Here $KL(\Theta_d, \Theta_{d_i}) = \sum_{t \in V} p(w|\Theta_d) \log \frac{p(w|\Theta_d)}{p(w|\Theta_{d_i})}$ and Jelinek-Mercer smoothing is used.

In order to calculate the *dynamicity property* of a text collection new documents of a collection are compared against the previous documents of the same collection. Documents added in the nearest past form a set of documents DS of size N , thus DS can be viewed as a window of size N . As a new document d arrives, its novelty score against DS is calculated. Then d is added to DS ,

while the oldest document in DS is removed (i.e. the window moves one step forward preserving only the last N documents). Thus the final dynamicity score is computed as follows.

```

Input: Collection  $C$  of documents sorted chronologically; set of
           documents  $DS$  formed out of first  $N$  documents in  $C$ 
Output: Sum of novelty scores of all documents in  $C$ 
foreach document  $d \in C \setminus DS$  do
     $DynScore(C) + = NS(d, DS)$ ;
     $DS = (DS \cup \{d\}) \setminus \{d_{oldest}\}$ ;
end
    
```

The final dynamicity score is normalized by the number of documents in a collection: $DynScore(C) = \frac{DynScore(C)}{|C|}$. Window size N is a parameter that can be chosen according to a collection statistics.

In order to measure the *diversity property* of a text collection we choose a random sample of documents RS of size M and compare these documents pairwise, since there is no notion of time in this case. Therefore the diversity score is calculated as follows: $DivScore(C) = \frac{\sum_{d \in RS} NS(d, RS \setminus \{d\})}{|RS|}$. Here M is a parameter that can be chosen according to a collection statistics.

4 Experiments

To test the proposed dynamicity and diversity measures we choose several intuitively different collections from TREC volumes 1-3, namely AP, WSJ, FR and Patents datasets. We expect the collections of news articles (AP and WSJ) to be ranked higher than FR and Patents according to both dynamicity and diversity properties, although it is not obvious which one of FR and Patents collections is more dynamic or more diverse.

The results for the dynamicity and diversity measures are presented in table II. Due to the space limitations only the dynamicity measures will be discussed. The values of the diversity measures are calculated in a similar way, therefore the following discussion is applicable to them as well.

The ranking of the collections according to their dynamicity property is the same for $N = 10$ and $N = 100$, therefore the proposed dynamicity measures

Table 1. Dynamicity and diversity measures

Coll.	Dynamicity						Diversity					
	AvgNWR		CosDistTF		AvgLM		AvgNWR		CosDistTF		AvgLM	
	$N = 10$	$N = 100$	$N = 10$	$N = 100$	$N = 10$	$N = 100$	$M = 100$	$M = 1000$	$M = 100$	$M = 1000$	$M = 100$	$M = 1000$
WSJ	0.84	0.85	0.85	0.86	2.18	2.71	0.85	0.85	0.85	0.86	2.74	2.91
AP	0.84	0.85	0.81	0.83	1.98	2.56	0.86	0.86	0.83	0.83	2.63	2.81
Patents	0.65	0.65	0.73	0.74	1.62	1.94	0.66	0.66	0.74	0.74	1.97	2.06
FR	0.67	0.72	0.41	0.45	1.41	1.89	0.73	0.72	0.45	0.45	1.92	1.98

are stable across different window sizes N . As we expected, all the measures clearly split the collections in question into two sets: more dynamic collections of news articles (AP and WSJ) and less dynamic FR and Patents collections. However, there is no clear ranking inside those sets: AvgNWR-based measure give higher dynamicity score to the AP collection compared to the WSJ and to the FR collection compared to the Patents. CosDistTF- and AvgLM-based measures, on the other hand, produce the reverse ranking: the WSJ collection is considered to be more dynamic than the AP and the Patents collection to be more dynamic than the FR. We are currently investigating these results.

5 Future Work

There are a number of possible applications of the proposed measures. In Distributed Information Retrieval resource description, resource selection and results fusion algorithms may benefit from the knowledge of the values of dynamicity and diversity properties of federated collections. Special smoothing techniques could be applied for collections known to be relatively homogeneous (i.e. not diverse). Also query expansion may use only terms from recent documents in relatively dynamic collections. We are going to address all these questions in future work.

References

1. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proc. of the ACM SIGIR. pp. 314–321. ACM (2003)
2. Callan, J.: Advances in Information Retrieval, chap. 5. Distributed Information Retrieval, pp. 127–150. Kluwer Academic Publishers (2000)
3. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of the ACM SIGIR. pp. 335–336. ACM (1998)
4. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proc. of the ACM SIGIR. pp. 275–281. ACM (1998)
5. Song, K., Tian, Y., Gao, W., Huang, T.: Diversifying the image retrieval results. In: Proc. of the ACM MM. pp. 707–710. ACM (2006)
6. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: Proc. of the ACM SIGIR. pp. 81–88. ACM (2002)
7. Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: Proc. of the WWW. pp. 22–32. ACM (2005)

Manuzio: A Model for Digital Annotated Text and Its Query/Programming Language

Marek Maurizio and Renzo Orsini

Dipartimento di Informatica
Universit Ca' Foscari di Venezia
Via Torino 155, Venezia Mestre, Italy
{marek,orsini}@dsi.unive.it

1 Introduction

More and more large repositories of texts which must be automatically processed represent their content through the use of descriptive markup languages. This method has been diffused by the availability of widely adopted standards like SGML and, later, XML, which made possible the definition of specific formats for many kinds of text, from literary texts (TEI) to web pages (XHTML). The markup approach has, however, several noteworthy shortcomings. First, we can encode easily only texts with a strict hierarchical structure while text has often concurrent hierarchies. Then, extra-textual information, like metadata or annotations, can be tied only to the same structure of the text and must be expressed as strings of the markup language. Third, queries and programs for the retrieval and processing of text must be expressed in terms of languages like XQuery [4], in which every document is represented as a tree of nodes; for this reason, in documents where parallel, overlapping structures exists, the complexity of XQuery programs becomes significantly higher.

Consider, for instance, a collection of classical lyrics, with the two parallel hierarchies lyric > stanzas > verses > words, and lyric > sentences > words, with title and information about the author for each lyric, and where the text is annotated both with commentary made by different scholars, and with grammatical categories in form of tree-structured data. Such a collection, if represented with markup techniques, would be very complex to create, manage and use, even with sophisticated tools, requiring the development of complex ad-hoc software.

To overcome some of the above limitations partial solutions exist (see for instance [3]), but at the expense of greatly increasing the complexity of the representation through difficult to read markup extensions, like the so-called “milestone” elements. Moreover, markup query languages need to be extended to take these solutions into consideration [1], making even more difficult to access and use such textual collections.

In the project “Musisque Deoque. A digital archive of Latin poetry, from its origins to the Italian Renaissance” sponsored by the Italian MIUR, we have built a model and a language to represent repositories of literary texts with any kind of structure, with multiple and scalable annotations, not limited to textual data,

and with a query component useful not only for the retrieval of information, but also for the construction of complex textual analysis applications. This approach fully departs from the markup principles, borrowing many ideas from the object-oriented models currently used in programming languages and database areas. A comprehensive description of the model, language, and system can be found in [5]. The language (called Manuzio) has been developed to be used in a multi-user system to store persistently digital collections of texts over which queries and programs are evaluated. This paper reports mainly the work done on the model and the language, since the system is still at its early stages of development with a prototypal implementation.

2 The Manuzio Model

The Manuzio model considers the textual information in a dual way: as a formatted sequence of characters, as well as a composition of logical structures called *textual objects*, similar to the content objects described in [2]. A *textual object* is a software entity with a state and a behavior. The state defines the precise portion of the text represented by the object, called the *underlying text*, and a set of *properties*, which are either *component* textual objects or *attributes* that can assume values of arbitrary complexity. The behavior is constituted by a collection of local procedures, called *methods*, which define computed properties or perform operations on the object. A textual object T is a *component* of a textual object T' if and only if the underlying text of T is a subtext of the underlying text of T' ¹.

The Manuzio model can also represent homogeneous aggregation of textual objects called *repeated textual objects*. Through repeated textual objects it is possible to represent complex collections like “all the first words of each poem” or “the words which form the subject in a certain sentence” as single textual objects. A *repeated textual object* is a set of textual objects of the same type, called its *elements*. Its underlying text is the composition of the underlying text of its elements.

Each textual object has a type, which represents a logical entity of the text, such as a word, a paragraph, a sentence, and so on. In the Manuzio model types are organized as a lattice where the greatest element represents the type of the whole collection, and the least is the type of the most basic objects of the schema. Types can also be defined by inheritance, like in object-oriented languages. For instance, the types Novel and Poem are both subtypes of Work. An example of textual schema is given, by the means of a graphical notation, in Figure 1. In the figure boxes represent textual object types, single and double pointed arrows represent component relationships with single or repeated textual object types, while hollow pointed arrows represent specialization relationships.

¹ Differently from a substring, a subtext can comprise non-contiguous parts of a text.

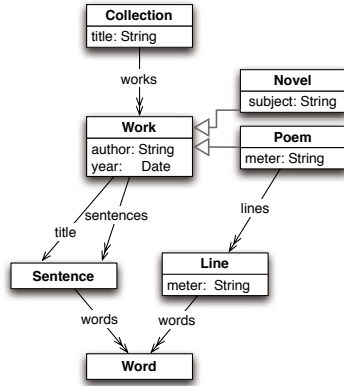


Fig. 1. Example of Manuzio Model

3 The Manuzio Language

Manuzio is a functional, type-safe programming language with specific constructs to interact with persistently stored textual objects. The language has a static and strong type system with which to describe schemas as that illustrated in Figure 1, and a set of operators which can retrieve textual objects without using any external query language. A persistent collection of documents can be imported in a program and its root element can be referenced by a special variable `collection` of type `Collection`. From this value all the textual objects present in the collection can be retrieved through operators that exploit their type’s structure: the *get* operator retrieve a specific component of an object, while the *all* operator retrieve recursively all the components and subcomponents of a certain type of an object. Other operators allow the creation of expressions similar to SQL or XQuery FLOWR expressions². Since the queries are an integrated part of the language, they are subject to type-checking and can be used in conjunction with all the other language’s features transparently.

The program in Source Code 1, for instance, assigns to a variable the first sentences of each work. This portion of text can be subsequently refined or used in any retrieval context. In Source Code 2 a more complex example is shown, where an analysis of Shakespeare’s plays extracts the top three “love speaking” characters in “A Midsummer Night’s Dream”. The results of such code are then reported in Source Code 3.

```
let incipits = select all SENTENCE 1 of works of collection;
```

Source Code 1. Retrieve the incipits of each work

² The full syntax and semantics of the Manuzio language can be found in [5].

```

let play = p in (get plays of collection) where p.title = "A Midsummer Night's Dream";
let loveSpeeches = s in (getall Speech of play)
    where some w in (getall Word of s) with (get stem of w) = "love";

let love_speech_count_by_speaker =
    select {speaker = s.speaker, n=(size of s.partition)}
    from s in (speeches groupby speaker);
output "The top 3 love speakers are:" + ove_speech_count_by_speaker[1..3];

```

Source Code 2. Compute a new structure of the most love-speaking characters

The top 3 love speakers are:

```

[{"speaker="LYSANDER", n=17},
 {"speaker="OBERON", n=13},
 {"speaker="HERMIA", n=12}]

```

Source Code 3. Results of Source Code 2

4 Conclusions and Future Work

To evaluate the usefulness of our approach a first prototype of the Manuzio language has been developed by mapping the textual objects into a relational database system and by testing it with small-sized corpora. We are aware that a great deal of work on data representation and query optimization must yet be done to provide a satisfying performance for large collections of texts. However, we think that work on modeling and linguistics aspects of retrieval of texts and computations over them is very important per se. In the near future we will explore new ways of providing the model operators, as, for instance, by writing libraries for well known languages, and we will start prototyping a multi-user system for supporting cooperative annotations on large text repositories.

References

1. Dekhtyar, A., Jacob, I.E., Kiernan, K., Porter, D.C.: Extended xquery for digital libraries. In: JCDL 2006: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, p. 378. ACM Press, New York (2006)
2. DeRose, S.J., Durand, D.G., Mylonas, E., Renear, A.H.: What is text, really? ACM SIGDOC Asterisk Journal of Computer Documentation 21(3), 1–24 (1997)
3. DeRose, S.J.: Markup overlap: A review and a horse. In: Extreme Markup Languages (2004)
4. Katz, H., Chamberlin, D.D.: XQuery from the experts: a guide to the W3C XML query language. Addison-Wesley, Reading (2004)
5. Maurizio, M.: Manuzio: an Object Language for Annotated Text Collections. PhD thesis, Dipartimento di Informatica, Università Ca' Foscari di Venezia (2010)

Effective Term Weighting for Sentence Retrieval

Saeedeh Momtazi¹, Matthew Lease², and Dietrich Klakow¹

¹ Spoken Language Systems, Saarland University, Germany

² School of Information, University of Texas at Austin, USA

Abstract. A well-known challenge of information retrieval is how to infer a user’s underlying information need when the input query consists of only a few keywords. Question Answering (QA) systems face an equally important but opposite challenge: given a verbose question, how can the system infer the relative importance of terms in order to differentiate the core information need from supporting context? We investigate three simple term-weighting schemes for such estimation within the language modeling retrieval paradigm [6]. While the three schemes described are ad hoc, they address a principled estimation problem underlying the standard word unigram model. We also show these schemes enable better estimation of a state-of-the-art class model based on term clustering [5]. Using a TREC QA dataset, we evaluate the three weighting schemes for both word and class models on the QA subtask of sentence retrieval. Our inverse sentence frequency weighting scheme achieves over 5% *absolute* improvement in mean-average precision for the standard word model and nearly 2% *absolute* improvement for the class model.

1 Introduction

Information Retrieval (IR) addresses a critical user need to be able to find relevant information in vast digital libraries. However, IR systems typically treat documents as the atomic unit of retrieval, and it is often the case that only a portion of any given document is actually relevant to the user’s information need. Another shortcoming of standard IR systems is their emphasis on putting the burden on the user to formulate short keyword queries. Such formulation becomes increasingly difficult as information needs become more complex and can often lead to iterative query reformulation and search abandonment.

In contrast to typical IR, Question Answering (QA) both supports focused retrieval and allow people to easily express their information needs as natural language questions. While it is a laudable goal to shift the burden of effort from users in query formulation to systems in query interpretation, a clear challenge lies in developing QA systems capable of effectively interpreting such queries. Addressing this challenge represents an important direction for long-term research, and this paper presents an early step toward this overarching goal.

A standard QA system architecture incorporates several subtasks: (i) retrieving relevant documents (ii) performing Sentence Retrieval (SR) from those documents, and (iii) extracting answers from sentences. In this work, we focus on improving accuracy of the SR component. In comparison to document retrieval,

SR poses several distinct challenges: (1) the brevity of sentences vs. documents exacerbates the usual term-mismatch problems, and (2) the verbosity of questions can lead to critical query terms being obscured by supporting terms. Various research has been done to improve SR performance, such as the ones proposed by Balasubramanian [2] and Allan [1]; however, to the best knowledge of the authors, there was no focus on the above problems. In this paper, regarding to (1), we build on recent work addressing term-mismatch via class-based modeling [5]. As for (2), we investigate three simple term-weighting strategies for approximating the relative importance of query terms: Inverse Document Frequency (IDF), Inverse Collection Frequency (ICF) [7], and a novel Inverse Sentence Frequency (ISF) scheme. While more elaborate and principled estimation schemes can be envisioned for inferring such relative importance, the above schemes are simple, efficient, and as results show, remarkably effective.

2 Method

In word unigram Language Model (LM) retrieval [6] and class model based on term clustering [5] sentences S are ranked by :

$$P_{word}(Q|S) = \prod_{q \in Q} P(q|S) \quad (1)$$

$$P_{class}(Q|S) = \prod_{q \in Q} P(q|C_q, S)P(C_q|S) \quad (2)$$

where $Q = \{q_1 \dots q_{|Q|}\}$ denotes a query of length $|Q|$. C_q is the cluster that contains q , $P(q|C_q, S)$ is the emission probability of q given its cluster and the sentence, and $P(C_q|S)$ is estimated based on clusters instead of terms.

Significant work has explored methods like Dirichlet-smoothing for better estimating the unigram models underlying observed documents, and the correlate here is the unigram model $P(Q|S) = \theta^S$ underlying S . Complimenting this work, KL-divergence ranking was described in which a latent unigram θ^Q is assumed to represent the user's information need underlying Q , and sentences are ranked via minimal KL-divergence between distributions [8]:

$$-D(\theta^Q || \theta^S) \stackrel{rank}{=} \theta^Q \cdot \theta^S \quad (3)$$

Of note here is that the standard LM approach is equivalent to KL-ranking only if θ^Q is estimated from Q via Maximum Likelihood (ML). Thus the standard unigram model can be understood as implicitly assuming a uniform distribution over query terms, meaning all query terms are inferred to be equally important to the underlying information need. While this assumption is reasonable for short keyword queries, it is increasingly problematic as query length increases due to large variance of relative term importance in natural language [4]. While not described in this way, Smucker and Allan [7] introduced ICF-weighting as a simple (and admittedly ad hoc) alternative to ML for better estimating θ^Q .

We present the first investigation of this strategy for the SR task. In addition to considering ICF, we also evaluate IDF and a similar ISF scheme which differs from IDF by counting sentences rather documents. We evaluate these three

schemes for estimating Θ^Q in the context of two methods for modeling Θ^S : directly (the standard word-based model [6]) and via the class-based approach described above. In all cases, Dirichlet-smoothing is used to estimate Θ^S .

3 Evaluation

We evaluate our SR models using questions from the TREC¹ 2006 QA track with the TREC 2005 set was used for development. Documents come from the AQUAINT corpus² of 450 million tokens of English newswire text. Because original TREC relevance judgments were only made at the coarser document level, we used the *Question Answer Sentence Pair* corpus of Kaisser and Lowe [3].

To evaluate the SR component of our QA system independent of the document retrieval component, we adopted the following experimental setup. A separate sentence collection was first created for each *question-series* (TREC QA data specifies each questions in the context of a series of related questions). For each series, we identify all documents known to be relevant to *any* question in the series, and we add all sentences from that document to the sentence collection. The average size of this collection is 270 sentences per question while the average number of relevant sentences per question is only 4. Moreover, the non-relevant sentences in each collection exemplify exactly the sort of typical QA system false alarms we want our SR system to avoid: non-relevant sentences coming from (1) documents relevant to similar yet different questions and (2) non-relevant sentences found in relevant documents.

IDF, ISF, and ICF statistics were taken from AQUAINT. We did the similar experiments with the larger Gigaword corpus³ and achieved similar trends which are not reported further. Following Momtazi and Klakow [5], we used the same approach to build the class model.

Table 1. Mean-average precision of different weighting methods for word and class models. * marks statistical significance at $p < 0.01$ for 2-tailed paired t -test.

Model	Baseline	Weighted		
		IDF	ICF	ISF
Word LM	0.3696	0.4211*	0.4096*	0.4244*
Class LM	0.4174	0.4233	0.4336*	0.4353*

Table 1 shows results for mean-average precision; similar improvements were seen with mean reciprocal rank. For both word and class models, ISF-weighting is seen to consistently perform best and yield significant improvement. While both ISF and IDF-weighting of the word model exceed accuracy of the baseline class model, IDF-weighting fails to improve the class model. While ICF and ISF-weighting yield similar improvements for the class model, ICF-weighting under-performs ISF-weighting for the word model.

¹ <http://trec.nist.gov>

² Linguistic Data Consortium corpus LDC2002T31

³ Linguistic Data Consortium corpus LDC2003T05

Comparing the results of ISF- and IDF-weighting, our intuition is that the more specific term contexts used by ISF is more informative. That is, ISF-weighting compiles its frequency statistics from narrower text segments in comparison to IDF-weighting, which uses a far wider context and does not consider the number of times a word appears in a specific context.

While the weighting schemes approximate term importance via simple frequency statistics, the class-based model clusters frequent terms into a single class which implicitly decreases their effect in a similar fashion. Thus it is not too surprising that the weighting schemes are somewhat less effective with the class-based model. Nevertheless, statistically significant improvement is still achieved over the baseline class-based model.

4 Summary

This paper showed a simple and effective way for integrating several alternative term weighting strategies with word or class models for sentence retrieval. While far more work will be needed to bring us closer to our long-term goal of supporting QA for rich, complex natural language questions, we believe this work represents a simple first step in this direction and provides a new, useful baseline to which more sophisticated methods can be later compared.

Acknowledgements. Saeedeh Momtazi is funded by the German research foundation DFG through the International Research Training Group (IRTG 715).

References

1. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proceedings of ACM SIGIR International Conference, pp. 314–321 (2003)
2. Balasubramanian, N., Allan, J., Croft, W.: A comparison of sentence retrieval techniques. In: Proceedings of ACM SIGIR International Conference, pp. 813–814 (2007)
3. Kaisser, M., Lowe, J.: Creating a research collection of question answer sentence pairs with Amazon’s mechanical turk. In: Proceedings of the LREC International Conference (2008)
4. Lease, M., Allan, J., Croft, W.B.: Regression Rank: Learning to Meet the Opportunity of Descriptive Queries. In: Proceedings of the European Conference on Information Retrieval (ECIR), pp. 90–101 (2009)
5. Momtazi, S., Klakow, D.: A word clustering approach for language model-based sentence retrieval in question answering systems. In: Proceedings of ACM CIKM International Conference, pp. 1911–1914 (2009)
6. Ponte, J., Croft, W.: A language modeling approach to information retrieval. In: Proceedings of ACM SIGIR International Conference, pp. 275–281 (1998)
7. Smucker, M., Allan, J.: Lightening the load of document smoothing for better language modeling retrieval. In: Proceedings of ACM SIGIR International Conference, pp. 699–700 (2006)
8. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)* 22(2), 214 (2004)

User-Oriented Evaluation of Color Descriptors for Web Image Retrieval

Otávio A.B. Penatti and Ricardo da S. Torres

Institute of Computing, University of Campinas (Unicamp),
Campinas, Brazil
{penatti,rtorres}@ic.unicamp.br

Abstract. This paper proposes a methodology for effectiveness evaluation in content-based image retrieval systems. The methodology is based on the opinion of real users. This paper also presents the results of using this methodology to evaluate color descriptors for Web image retrieval. The experiments were performed using a database containing more than 230 thousand heterogeneous images that represents the existing content on the Web.

Keywords: user evaluation, color descriptors, content-based image retrieval, web.

1 Introduction

The growth in the size of image collections and the availability of these collections worldwide is an evident trend. The Web is, nowadays, one of the biggest and most heterogeneous image databases existent. The existence of this huge amount of visual information has increased the demand for image retrieval systems. A promising approach to address this demand is to retrieve images based on image content (Content-Based Image Retrieval - CBIR). This approach considers image visual properties for indexing and retrieval. One of the main visual properties considered by human vision in images is color. In CBIR systems, color is the most common property analyzed and it is the most studied in the literature [1].

A CBIR system is based on *image descriptors*. The image descriptor is composed by [2]: (i) an algorithm for extracting feature vectors and (ii) a distance function. The feature vector encodes information about image properties. Given two feature vectors, the descriptors' distance function computes a distance value. Given a query image, the distance value is used to rank database images.

The query processing time and the results of a search in a CBIR system depends on the descriptors used. Therefore, different descriptors can be used to achieve the system goals. This illustrates the importance of evaluating image descriptors considering different criteria. One important criterium is the descriptor effectiveness, which refers to the quality of the retrieved images.

In the literature, descriptors' effectiveness evaluation is usually made automatically. This evaluation is easier to be made, however it is possible to be

performed only in classified image databases. In a Web environment it is impossible to know the exact number of relevant images given a query image. Another problem with automatic effectiveness evaluation is that the previous database classification assumes that all users agree with the classification made. However, different users tend to have different interpretations about the same image and a single classification cannot be considered as true for all users.

An alternative to solve these problems is to evaluate descriptors considering the opinion of real users. One of the main advantages of a user-oriented evaluation is that the descriptors' effectiveness is measured by the opinion of potential users, what reflects a real environment of use for a CBIR system. Despite the large number of color descriptors in the literature, they are rarely tested in really heterogeneous environments like the Web and using potential users.

This paper performs experiments using a heterogeneous image database containing more than 230 thousand images collected from the Web. The main contributions are: (i) a methodology for effectiveness evaluation for CBIR systems by real users and (ii) the effectiveness evaluation of color descriptors for Web image retrieval.

2 A Methodology for User-Oriented Effectiveness Evaluation

The evaluation of image descriptors for CBIR tasks is very important and can be made based on different criteria. Important criteria include the complexity of feature extraction algorithms and distance functions, the storage requirements and the effectiveness of the descriptors. A theoretical comparative study of color descriptors for Web image retrieval is presented in [1].

In this paper we focus on the effectiveness evaluation. Effectiveness measures the descriptor's ability to retrieve relevant images. A descriptor with good effectiveness retrieves the most similar images at the first positions of the results (ranked list), for a given query image. The success of a CBIR systems is closely related to the quality of their results. A user can tolerate a not so fast response, but he or she will not tolerate non-relevant results.

In [3] a user-oriented evaluation is used for digital libraries. The effectiveness of structured and non-structured queries in digital libraries is evaluated by real users. A similar evaluation is proposed here to assess the effectiveness of image descriptors. A Web interface shows to the user the query image and a set of retrieved images. Users are asked to indicate which of that images they consider similar to the query image.

The set of images showed to the user is created as follows. For each descriptor, a list containing the 30 most relevant images is obtained for a given query image. The lists of each descriptor are combined, eliminating duplicates. Every image in the list has a reference to the descriptor(s) that retrieved it and to the rank in its original list. The final list is shuffled and showed to the user.

The query image is showed highlighted at the top of the page and the shuffled list is showed below it. The user has no information about which descriptor

retrieved each image. After the user has indicated the images they considered similar to the query image, they click on a button to finish the evaluation of that query image. The process repeats until all query images are evaluated.

When the user finishes the evaluation, effectiveness measures are computed. The measures used here are: P_{10} , P_{20} , and P_{30} . These measures stand for precision values and they indicate the percentage of images marked as similar among the top 10, 20, and 30 results, respectively. The measures are computed for each descriptor, for each query image and for each user. Therefore, it is possible to compute the descriptors' effectiveness for each query image independently.

3 Experiments and Results

The objective of the experiments was to evaluate color descriptors for Web image retrieval. The image database used was collected by researchers from Federal University of Amazonas (UFAM), Brazil. The database collection was made recursively from addresses found in the Yahoo directory. The final collection contains 234.828 images and more than 1 million of HTML documents. The image database has no classification.

Five color descriptors were evaluated in this paper: *global color histogram* (GCH) [4], *color autocorrelogram* (ACC) [5], *color structure* (CSD) [6], *border/interior pixel classification* (BIC) [7], and *color bitmap* [8]. They were chosen because of their simple algorithms for feature extraction and distance computation. Also, GCH, ACC, and CSD are important descriptors from the literature.

The effectiveness evaluation used 16 query images. These images have a well defined semantics and represent different image categories. 69 subjects were invited to take part in our experiments. Users that have not evaluated all the 16 queries were discarded. Therefore, the complete evaluation considered 15 users.

Table 1 presents the average values of P_{10} , P_{20} , and P_{30} among all query images for each descriptor. The results indicate that BIC descriptor has the best average precision for 10, 20, and 30 retrieved images. ACC and GCH achieved the second and third best average precision, respectively. CSD and Color Bitmap achieved similar average precision values. Although BIC descriptor has presented the best average precision, it was not the best for all query images [4].

Table 1. Average values of P_{10} , P_{20} , and P_{30} among all query images

Descriptor	P_{10}	Descriptor	P_{20}	Descriptor	P_{30}
BIC	0.31	BIC	0.21	BIC	0.17
ACC	0.27	ACC	0.18	ACC	0.15
GCH	0.25	GCH	0.17	GCH	0.13
CSD	0.18	Color Bitmap	0.12	Color Bitmap	0.10
Color Bitmap	0.17	CSD	0.12	CSD	0.09

¹ The precision tables for each query image are available at: http://www.lis.ic.unicamp.br/~otavio/cbir_eval/tables.pdf

Considering a Web scenario, it is very important that the relevant images appear at the top positions of the results. This indicates that it is more important to have higher values of P_{10} than P_{30} .

It is possible to note that the precision values decreased as the number of images increased. This indicates that when more images are retrieved, the descriptors retrieve more non-relevant images than relevant images.

The analysis of the precision values for each of the query images led us to some conclusions. Descriptors have presented higher precision values for queries with homogeneous background. Another aspect was the influence of the image semantics during the evaluation. Users tend to consider similar images with similar semantics and not necessarily with similar visual properties.

4 Conclusions

This paper presented a methodology for user-oriented effectiveness evaluation and the results of the evaluation of color descriptors for Web image retrieval. The experiments were performed in a heterogeneous image database with more than 230 thousand images collected from the Web. The results indicate that good color descriptors are BIC and ACC. Future work includes the evaluation of more descriptors and the inclusion of more users in the evaluation process.

Acknowledgments. Thanks to Fapesp (process numbers 2006/59525-1 and 2009/10554-8), Capes, CNPq, and Microsoft Research for financial support.

References

1. Penatti, O.A.B., da S. Torres, R.: Color descriptors for web image retrieval: A comparative study. In: XXI Sibgrapi, pp. 163–170 (2008)
2. da S. Torres, R., Falcão, A.X.: Content-Based Image Retrieval: Theory and Applications. *Revista de Informática Teórica e Aplicada* 13(2), 161–185 (2006)
3. Gonçalves, M.A., Fox, E.A., Krowne, A., Calado, P., Laender, A.H.F., da Silva, A.S., Ribeiro-Neto, B.A.: The effectiveness of automatically structured queries in digital libraries. In: JCDL, pp. 98–107 (2004)
4. Swain, M.J., Ballard, D.H.: Color indexing. *International Journal of Computer Vision* 7(1), 11–32 (1991)
5. Huang, J., Kumar, S.R., Mitra, M., Zhu, W., Zabih, R.: Image indexing using color correlograms. In: CVPR, p. 762 (1997)
6. Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Tech.* 11(6), 703–715 (2001)
7. de Oliveira Stehling, R., Nascimento, M.A., Falcão, A.X.: A compact and efficient image retrieval approach based on border/interior pixel classification. In: CIKM, pp. 102–109 (2002)
8. Lu, T., Chang, C.: Color image retrieval technique based on color features and image bitmap. *Information Processing and Management* 43(2), 461–472 (2007)

A Topic-Specific Web Search System Focusing on Quality Pages

Ari Pirkola and Tuomas Talvensaari

Department of Information Studies and Interactive Media, University of Tampere, Finland
{ari.pirkola,tuomas.talvensaari}@uta.fi

Abstract. We describe a topic-specific Web search system focused on quality pages and argue that there is a need for such quality-based topic-specific search tools. The first implementation of the search system is available on the Web and it deals with climate change. The key idea is to crawl (using a focused crawling technique) in known trusted sites and in sites that are connected to them. We also discuss the further development of the system and our future research. Our project plan involves building a larger quality-based Web search system dealing with many globally significant topics (in addition to climate change).

Keywords: Digital libraries, Focused crawling, Vertical search engines, Web information retrieval.

1 Introduction

We describe in this paper a topic-specific Web search system focusing on *quality documents* (i.e., pages) on climate change. The main advantage of the system is that users are not discouraged with low quality documents because documents for the system are crawled from known trusted sites and from the sites that are linked to them. In our approach, *quality site* refers, first, to the Web site of a university (in any country around the world). Information published on universities' Web sites is generally expected to be correct and reliable, based on facts and scientific findings and the standards of scientific ethics. This justifies to consider documents published on universities' Web sites as quality documents and the sites as quality sites. Second, a site pointed to by a link within a document of a university site is considered as a quality site. That is to say, in this approach a link within a university document pointing to another document located outside the network of universities' Web sites means that the referred document, and the whole site, is considered as a quality document / site. Universities' authors are a kind of a "cognitive filter": they are assumed to link their documents to quality documents, not to documents of poor quality.

The most popular search engines that provide access to Web documents are Google, Microsoft's Bing, and Yahoo. Their repositories contain billions of documents. Even though these major search engines are useful tools and often provide users with good results, a closer look suggests that there would be a better way to serve Web users with information needs related to specific domains or topics. In the major search engines, a query typically matches from thousands to millions of

documents, and search results are often very long. Therefore they need to be ranked, which is often based on the popularity of pages, as in Google's PageRank algorithm [1] which rewards documents that are pointed to by documents that themselves are popular documents. The problem is that it is often difficult to identify in search results documents that at the same time are *relevant* and *of high quality* [2, 3, 4]. Popular documents of poor quality may appear at the top of search results lists. Obviously such documents are of no use for most searchers, but they can even be harmful, in particular in the field of health and medicine. Typically, the pages of commercial organizations populate the top ranks while scientifically-oriented pages shine in their absence. On the other hand, reliable but unpopular documents containing highly relevant information that would best fulfil the user's information need are often ranked low in the search results.

2 CliCS System

Documents for our search system, which is called *CliCS - Climate Change Search* - were retrieved from the Web using focused crawling. A *focused crawler* is a program aiming to fetch Web documents that are relevant to a pre-defined domain or topic [5, 6, 7]. A large set of URLs of universities and trusted sites associated with universities were used as start URLs in crawling. The start URLs were received from the Webometrics World University Ranking [8] which is a well-known university ranking system. For each URL, the scope of crawling was restricted, so that only documents contained in universities' Web sites and their immediate child documents contained in any Web sites were selected for the CliCS system.

In crawling we used the *Nalanda iVia Focused Crawler* [9] that was modified so that, instead of using a text classifier of the original Nalanda, relevance scores to pages were assigned by matching a topic-defining query against the retrieved page. Here the Lemur search engine [10] was used as the query engine. The modified crawler is described in [6].

The following query was used to represent the topic climate change: *#wsum(2.0 #3(climate change) 2.0 #3(global warming) 2.0 #3(climatic change) 2.0 #3(climate research) 1.0 #3(research project) 1.0 research)*, where *#wsum* is the weighted sum operator, and *#3* the proximity operator, which means that the enclosed query keys are not allowed to be more than three words apart from each other to match. The query gives more weight to phrases that relate directly to global warming or climate change, and lower weight to keys and phrases that relate to research activity in general.

The crawled relevant pages were indexed using the Apache Lucene [11] programming library. The search engine is also powered by Lucene. Lucene supports ranked boolean queries, as well as wildcard queries (e.g., *climat* change*), and queries targeted at specific fields (e.g., title) of the indexed documents. Currently, CLiCS has indexed 15 891 documents. The current implementation is available at <http://kastanja.uta.fi:8988/CLICS/>.

A thorough assessment of the *accuracy* of the documents indexed by a retrieval system would require a panel of domain experts to assign "accuracy scores" to the documents. To get reliable assessments, the experts should assess a large number of

documents returned by the system in response to queries in that domain. This type of system evaluation would be a huge effort and is beyond a reasonable effort in a typical scientific study.

However, to get a view of the quality of the document collection indexed by a retrieval system the retrieval results need to be evaluated or described in some manner. For this purpose, we queried the CliCS system and categorized the retrieved documents (top 20 documents for each query) based on the sites in which the documents were contained. The following five terms were selected as queries from the Wikipedia article discussing global warming and were run in the CliCS system: *greenhouse gas*, *solar radiation*, *melting*, *fossil fuels*, and *methane*. The categories and the numbers of retrieved documents are shown in Table 1. *University* (the first row) refers to a university or its department, or a centre, institute, or a school associated with a university. As was expected, academic sites dominate in CliCS's results. There are only four commercial sites. As documents published on universities' Web sites are expected to be reliable, following the standards of scientific practice, we can conclude that the CliCS system returns scientifically-oriented quality documents.

Table 1. Site types in CliCS results for the five queries

Site type	Number of retrieved documents
A. University	65
B. Professional, scientific, or research org. other than A	16
C. Non-commercial organization other than A or B	6
D. Commercial organization	4
E. Government agency	7
F. Other	2
All	100

3 Discussion and Conclusions

We described the first implementation of our topic-specific Web search system focusing on quality documents in the research area of climate change. The quality-based crawling approach is original involving a new key idea: crawling in trusted sites and "cognitive filtering" based on the links in the pages of the trusted sites. This idea should be developed further. In this study, we selected from the results of crawling only documents that were one link apart from a university site. However, it may be that longer link chains are as good as one link. When crawling is started from the university site, at some point it reaches an end point where relevant documents are not found any more. Is it so that document quality remains high when a link chain is followed from the university site to the end point? Or is it so that at some point quality starts to degrade? If the quality does not degrade, crawling through longer link chains would allow a means to increase the coverage of a quality-based search system. It is the task of future research to shed light on these questions.

The further development of the CliCS system involves adding an uploading option (and an associated quality filter) as well as adding multilingual features to the system. Uploading means that authors and publishers can themselves submit documents and URLs to the system. Therefore, the transfer of documents and URLs to the system will be a collaborative effort. We expect that authors will submit in particular such documents with which general search engines have difficulty: documents that only appear in the Deep Web and new documents in the surface Web.

Adding multilingual features to the system means that documents in other major languages than English will be crawled and made accessible through the system. Even though English is the dominant language of science, a large number of scientific works and scientifically-oriented pages published on the Web are written in languages other than English. It is therefore important to develop multilingual focused crawling methods and multilingual topic-specific search systems.

Our project plan involves building a multi-topic Web search system, so that the CliCS system will be a part of a larger quality-based system, dealing with many globally significant topics (in addition to climate change). We believe that such a multi-topic quality-based system best serves users, such as researchers and journalists searching for information on scientifically and globally important topics.

Acknowledgments. This study was funded by the Academy of Finland (research projects 129835, 130760, 218289).

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1-7), 107–117 (1998)
2. Widiantoro, D.: Toward the development of next generation search engine. In: *International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia (2007)
3. Griffiths, K., Christensen, H.: The quality and accessibility of Australian depression sites on the World Wide Web. *Medical Journal of Australia* 176, S97–S104 (2002)
4. Krones, C., Böhm, G., Ruhl, K., Stumpf, M., Klinge, U., Schumpelick, V.: Inguinal hernia on the Internet: A critical comparison of Germany and the U.K. *Hernia* 8(1), 47–52 (2004)
5. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. In: *Eighth International World Wide Web Conference*, Toronto, May 11-14 (1999)
6. Pirkola, A., Talvensaari, T.: Addressing the limited scope problem of focused crawling using a result merging approach. In: *25th Annual ACM Symposium on Applied Computing (ACM SAC)*, Sierre, Switzerland, March 22 - 26, pp. 1735–1740 (2010)
7. Tang, T., Hawking, D., Craswell, N., Griffiths, K.: Focused crawling for both topical relevance and quality of medical information. In: *Fourteenth ACM International Conference on Information and Knowledge Management, CIKM 2005* (2005)
8. Webometrics University Ranking, <http://www.webometrics.info/>
9. Nalanda iVia Focused Crawler, <http://ivia.ucr.edu/>
10. Lemur search engine, <http://www.lemurproject.org/>
11. Apache Lucene, <http://lucene.apache.org/>

Reliable Preservation of Interactive Environments and Workflows

Klaus Rechert, Dirk von Suchodoletz, Randolph Welte,
Felix Ruzzoli, and Isgandar Valizada

Albert-Ludwigs University Freiburg, Hermann-Herder Str. 10,
79104 Freiburg i. B., Germany

Abstract. The creation of most digital objects occurs solely in interactive graphical user interfaces which were available at a particular time period. Archiving and preservation organizations are posed with large amounts of such objects of various types. At some point they will need to automatically process these to make them available to their users or convert them to a commonly used format. We present methods and a system architecture for emulation services which enable the preservation of interactive environments and their workflows in a reliable manner. This system includes a framework for describing interactions with an interactive environment in an abstract manner, for supporting reliable playback in an automated way and finally for ensuring the preservation of specific operation knowledge by documenting and storing all components in a dedicated software archive.

1 Introduction

Long-term preservation and accessibility of digital objects pose new and diverse requirements. The creation of most digital objects has occurred solely in interactive graphical user interfaces which were available at a particular time period. Archiving and preservation organizations have already accumulated a large quantity of such objects of various types. A substantial challenge is to allow migration of digital objects within their original application in an automated and controlled way. Availability of suitable tools poses a major problem for this task.

2 Automation of Interactive Workflows

Typical digital objects in memory institutions were created with interactive applications on computer architectures with graphical user interfaces. The user was required to point and click or use the keyboard to create or modify an object. Migration steps would require mostly the same type of actions to be executed. If larger quantities of objects are to be handled, a manual procedure would be time consuming, expensive and error-prone. The traditional approach to help the user to automate interactive tasks to a certain degree is the use of so-called

macro-recorders. These are specialized tools or functions of an application or operating system user interface to capture sequences of executed actions. However, this functionality is not standardized in terms of its usability and features. Special software components are needed, but also knowledge of the applications and operating systems is necessary. A generic approach demands a technical and organizational separation between the machine used for executing workflows and its input/output. Hence, emulated or virtualized environments are particularly well suited for recording an interactive workflow once, such as installing a specific printer driver for PDF output, loading an old Word Perfect document in its original environment and converting it by printing into a PDF file. Such a recording can also serve as the base for a deeper analysis and the generation of a machine script for the future. By using the aforementioned method, the authors demonstrated the feasibility of such simple migration task in an automated way [12].

2.1 Improving Reliability of Replaying Recorded Interactions

An interactive workflow can be described as an ordered list of interaction events. Interactions might be mouse movements or keystrokes and are passed on to the emulated environment through a defined interface at a particular time. By using a generic approach to describe interactive events, there is usually no explicit feedback to executed interactive events. While a traditional macro-recorder has good knowledge of its runtime environment (e.g. the capability to communicate with the operating system) in a generic emulation setup usually only the screen output and the internal state of the emulated hardware is visible (e.g. CPU state, memory). Furthermore, the recording/playback system has no knowledge of the system it operates. Thus, a framework for playing back a complete workflow in a reliable way is indispensable. A solution relying solely on the time elapsed between the recorded action is not sufficient: executing recorded actions will take different amounts of time to complete depending on the load of the host-machine and the state of the runtime environment. Therefore, we link each interaction with a precondition and an expected outcome which can be observed as a state of the emulated environment. Until this effect is observed, the current event execution has not completed successfully and the next event cannot be processed. While, in case of human operation, the effect is observed through visual control in an automated run an abstract definition of expected states and their reliable verification is indispensable. One suggested solution makes use of visual synchronization points [3]. For example, a snapshot of a small area around the mouse cursor can be captured before and after a mouse event and then used for comparison at replay time. Hence, replaying an interactive workflow becomes independent of computation time the host-machine needs to complete a particular action execution. However, removing time constraints still cannot guarantee a reliable playback in general. First, if the synchronization snapshot is done in an automated way, important aspects of the observable feedback on executed actions might get lost. An optional manual selection of the snapshot area can improve the reliability since the user is carrying out the recording and is usually

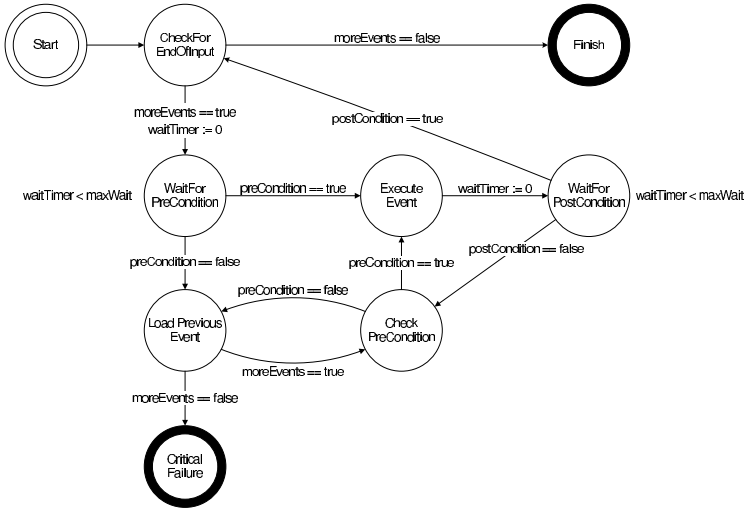


Fig. 1. Abstract state machine to execute a sequence of interactive events

familiar with the interaction model of the graphical environment he operates. Second, mouse and keyboard events are passed on to the runtime environment through an abstract interface (e.g. through hardware emulation of a PS/2 mouse interface). Hence, sometimes the environment does not react to input events in the expected way. This occurs for example if the operating system is busy and unable to process input events. For reliable playback such failures need to be detected and handled by the framework. Furthermore, the operator needs support to implement specific failure recovery strategies, e.g. resetting the machines to a stable previous state and retry the failed subsequence. If the operator is additionally able to attach meta data to specific events describing its original intend and possible side effects, not only the reliability of automated execution will be improved but also specific knowledge of practical operation will be preserved.

To support these ideas, the interactive workflows has to be represented as time independent event transitions, only relying on valid stable pre- and postconditions. For describing pre- and postconditions the visual snapshot from VNCPlay [3] was used, but extended to support users to choose the relevant snapshot area. A timeout value ensures termination of any given workflow. The framework accepts three types of input events: keyboard entry, mouse events and special pseudo-events. Pseudo-events include specific control commands of the runtime environment (e.g. ctrl-alt-del) but might also be used to map the progress of longer running tasks (e.g. installation procedure) through empty dummy events. Since the abstract event-passing interface provides no guarantees on action execution, especially a mouse pointer placement and verification system had to be implemented. Such a system does not only make mouse movement independent of the original users movements, but also allows to jump to any previous event with a defined mouse pointer state.

Figure 10 describes the execution of a given interactive action sequence. State transitions are triggered either through the arrival of appropriate feedback from the runtime environment or through a timeout. Failures can happen either by mismatching the precondition or the postcondition. If the precondition is not met within a defined timeout, the system could try to step back until a previous precondition matches and could retry event execution from that point. In case of a mismatched postcondition, the system could check if the precondition still holds and retry the last event execution. Although both recovery strategies may cover the most common failures, the operator still needs to decide which strategy is appropriate.

3 Conclusion and Outlook

The system presented includes a framework for describing interactions with an interactive environment in an abstract manner, supporting reliable playback in an automated way and finally ensure the preservation of specific operation knowledge by documenting and storing all components in a dedicated software archive. VNC offers an appropriate abstract layer to operate a standard computer interface providing screen output and keyboard and mouse input. Thus, it seems to be desirable to equip hardware emulators of different types with such an additional interface as found on QEMU used in our experiments. The procedures presented could be applied to completely different GUIs of other operating systems. Additionally monitoring the emulated machines internal state could improve reliability and predictability further. This way it becomes possible to detect system crashes or machine overload situations.

Acknowledgments

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

References

1. Rechert, K., von Suchodoletz, D., Welte, R., van den Dobbeltstein, M., Roberts, B., van der Hoeven, J., Schroder, J.: Novel workflows for abstract handling of complex interaction processes in digital preservation. In: Proceedings of the Sixth International Conference on Preservation of Digital Objects, iPRES 2009 (2009)
2. Rechert, K., von Suchodoletz, D.: Tackling the problem of complex interaction processes in emulation and migration strategies. *ERCIM News* (80), 22–23 (2010)
3. Zeldovich, N., Chandra, R.: Interactive performance measurement with vncplay. In: ATEC 2005: Proceedings of the Annual Conference on USENIX Annual Technical Conference, pp. 54–64. USENIX Association, Berkeley (2005)

Automated Country Name Disambiguation for Code Set Alignment

Gramm Richardson

U.S. Department of Defense
gpricha@tycho.ncsc.mil

Abstract. Multiple standards and encodings for names of countries, as well as multiple renderings of the country names themselves cause problems for interoperability. This impacts both human and automated processing. This paper describes an automated method for aligning pairs of country code sets by examining the string similarity between the names of the countries in each set.

1 Introduction

Many schemes exist to define short alphanumeric representations for countries. Because of incomplete or non-existent mappings between these representations, valuable information can be lost. In fields from aviation or automobile manufacture to banking, telecommunications, and government applications, dozens of these code sets are in use and treated as authoritative for the particular sectors they cover. In different code sets, the code itself may refer to different sorts of entities. It may refer to a sovereign nation, a dependent territory, a group of territories, or a subdivision of another entity in the code set.

Because the number of geopolitical entities in a particular code set or standard is generally small (around 150-250), hand alignment is possible for small numbers of code sets. But to align many code sets, an automated system can save substantial time and effort. By matching country names to other similar names that refer to the same country, country code sets can be mapped and aligned in a relatively complete and accurate way, requiring a minimum of human intervention to verify the matches.

1.1 Choice and Structure of Code Sets

To test the soundness of this approach for aligning potentially dozens of country code sets, a standard test set of four country code sets was used – FIPS 10-4, ISO 3166-1, USAID ADS 260, and ITU-T x.121. The relation between code and country varies from code set to code set. While ITU-T x.121 assigns multiple codes to some countries, another ITU recommendation, e.164, assigns the international dialing code “1” to 24 entities. With these sorts of variations, a country code cannot be assumed to be a unique identifier for a country.

2 Process

Beginning with two sets of countries x and y , for every pair of countries in the Cartesian product of the sets, a similarity measure is calculated. To filter the results, all but the highest rated match for each country is discarded. Then only countries who have each other as a mutual highest rated match are returned. As a result of the filtering, from the original $\|x\| \times \|y\|$ comparisons, only the most likely pairs of matching names are returned to the user for final validation.

Countries are compared on the basis of their (multiple) names. So the similarity above can be considered a “level-two” similarity [2]. The proposed match assessed between the countries will be equal to the highest-rated matching pair of their names. The more names that a country can be identified by, the higher the chance of matching it with another country. The order in which country names or country codesets are compared does not affect the results. Generally, two names are more likely to refer to the same country the more similar the names are.

2.1 Precision and Recall

The reported precision of the match between a pair of code sets is the probability that any particular proposed matching pair of countries is correct. Because the countries represented in each code set are not known beforehand, the number of expected matches is estimated rather than known. In estimating recall, the number of matches between two code sets of lengths m and n is evaluated to be equal to the length of the smaller code set. In practice, this returns a lower-than-actual recall.

2.2 String Normalization

To minimize the rendering differences in names, several string normalizations are applied. The characters in the country names are normalized on the basis of letter case, punctuation, and spacing. The characters are converted to the Basic Latin Unicode range (e.g. ‘ô’ becomes ‘o’). Common words such as “of” and “republic” are removed from country names.

Systems like Getty’s Synoname [7] hard code stop words into their system. Whereas here, words that occur four or more times in the two lists of country names being compared are deemed not to be good indicators of a unique country name and are automatically removed. From a strict string-similarity standpoint, there is no reliable way to automatically associate “North Korea” with “Democratic People’s Republic of Korea”, rather than “Republic of Korea”, as each of the words in these country names meets or exceeds the four-occurrence threshold and are removed. To disambiguate such countries, real-world knowledge is required. A final normalization technique is to alphabetize the tokens in the country name string. This serves a similar function to the “bag of tokens” approach [1] in converting strings to vectors, allowing for a greater matching between country names with different word orders.

String normalization, particularly in the automated removal of commonly-occurring words, produces a median 0.263 improvement in F-Score ($\Delta p + 0.065$, $\Delta r + 0.034$) over non-normalized strings for every algorithm.

3 String Matching Algorithms

Several string similarity measures were compared in order to determine the comparison metric that produced the best mappings between code sets — Jaccard Index [8], Dice’s Coefficient [4], Damerau-Levenshtein Edit Distance [5], Jaro Distance Metric [6], Longest Common Substring, Longest Common Subsequence [4]. Every country name in one set was compared to every country name in another set. String similarity is normalized to a value between 0 and 1. Ignoring this normalization can result in measurements that cannot be directly compared [8]. In each case, 0 indicates no similarity between strings, and 1 indicates an exact match.

While French et al. [3] successfully clustered variants of academic institution names with a hybrid approach using edit distance as a character-level measure and the Jaccard Index as a word-level measure, the process in this paper compares country names on a character or n-gram level.

The n-gram counts of the country name strings in the two sets being compared can inform Dice’s Coefficient, similar to comparing vector representations of strings using TF-IDF and cosine similarity. Weighting Dice’s Coefficient (trigrams) by n-gram frequency gives a 0.013 increase in F-score ($\Delta p + 0.006$, $\Delta r + 0.03$) compared to the string normalization described previously. Compared to no weighting and no normalization, the improvement in F-score from applying n-gram frequency weighting is 0.158 ($\Delta p + 0.001$, $\Delta r + 0.26$).

4 Filtering Mechanisms

After generating a similarity measure between every country in two code sets, the results need to be filtered before being displayed for final confirmation. Matches rated at 0.0 are discarded. Rather than setting a threshold, higher precision and recall is achieved by suggesting countries sharing a mutual best match. If country *A* has country *B* as its best match, and *B* has *A* as its best match, then the match is accepted. However, if *A* has *B* as its best match, but *B* has any other country as its best match, then *A* is concluded to have no match. Compared to a 99% threshold (accepting any match rated higher than 99% of the matches in a given pair of code sets), F-Score when using mutual best matching increases from 0.7946 to 0.9429 ($\Delta p + 0.276$, $\Delta r - 0.016$).

5 Conclusion and Follow-On Work

The precision and recall results are likely reaching the limits of what is possible with non-intelligent string matching. Further improvements will require

Table 1. Results for country code set alignment applying optimal techniques

F-Score (p, r)	FIPS 10-4	ITU-T x.121	USAID ADS 260
ISO 3166-1	0.966 (0.969, 0.963)	0.981 (0.991, 0.972)	0.895 (0.926, 0.866)
FIPS 10-4	–	0.962 (0.975, 0.95)	0.981 (0.981, 0.981)
ITU-T x.121	0.962 (0.975, 0.95)	–	0.944 (0.953, 0.936)

knowledge of relationships between countries and country names to be added to the system. The following chart shows the F-Score, precision, and recall of the matches between pairs of code sets.

This approach could also be used to find matches in two lists of similar named entities (perhaps organizations or personal names), either to match named entities when there is a slight variation in names, or to determine how strong is the correlation between the lists.

This system proves to be a highly accurate method of aligning two sets of country names. As an aid to manual alignment, it can reduce a modeler's work from aligning hundreds of names, down to validating only a few. Automated assistance could also significantly shorten the length of time required to produce a country ontology to record the current and historical names of countries and assist in knowledge base population.

References

1. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: KDD 2003: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 39–48. ACM, New York (2003)
2. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web, pp. 73–78 (2003)
3. French, J.C., Powell, A.L., Schulman, E.: Using clustering strategies for creating authority files. *Journal of the American Society for Information Science* 51(8), 774–786 (2000)
4. Kondrak, G.: N-gram similarity and distance. In: Consens, M.P., Navarro, G. (eds.) SPIRE 2005. LNCS, vol. 3772, pp. 115–126. Springer, Heidelberg (2005)
5. Navarro, G.: A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)* 33(1), 31–88 (2001)
6. Piskorski, J., Sydow, M.: String distance metrics for reference matching and search query correction. In: 10th Business Information Systems Conference, pp. 353–365 (2007)
7. Siegfried, S.L., Bernstein, J.: Synoname: Getty's new approach to pattern matching for personal names. *Computers and the Humanities* 25(4), 211–226 (1991)
8. van Rijsbergen, C.J.: *Information Retrieval*, 2nd edn. Butterworths, London (1979)

LIFE-SHARE Project: Developing a Digitisation Strategy Toolkit

Beccy Shipman¹, Matthew Herring², Ned Potter¹, and Bo Middleton¹

¹ Health Sciences Library, University of Leeds, Leeds, LS2 9JT, UK

² J.B.Morrell Library, University of York, York, YO10 5DD, UK
{b.shipman,e.potter,m.m.middleton}@leeds.ac.uk,
mh43@york.ac.uk

Abstract. This poster will outline the Digitisation Strategy Toolkit created as part of the LIFE-SHARE project. The toolkit is based on the lifecycle model created by the LIFE project and explores the creation, acquisition, ingest, preservation (bit-stream and content) and access requirements for a digitisation strategy. This covers the policies and infrastructure required in libraries to establish successful practices. The toolkit also provides both internal and external resources to support the service. This poster will illustrate how the toolkit works effectively to support digitisation with examples from three case studies at the Universities of Leeds, Sheffield and York.

Keywords: digitisation, digital lifecycle, toolkit, strategies, libraries.

1 Introduction

The LIFE-SHARE project received funding from JISC in September 2009 to explore digitisation across the three White Rose Universities of Leeds, Sheffield and York. The project runs from September 2009 to February 2011. The aim of the project is to identify and establish institutional and consortial strategies and infrastructure for the creation, curation and preservation of a variety of digital content.

Digitisation is a complex procedure, or rather a complex network of parallel procedures. The work of the LIFE-SHARE Project is based on all aspects of the digital content lifecycle. The Project has adopted one of the current lifecycle models for digital content from the LIFE project [1] and uses this to analyse current practices within institutions and across the consortium. Each of the partner libraries is engaged in digital content creation, for a variety of purposes, and there is a growing need to understand costs and benefits across the digital content lifecycle. The adoption of the lifecycle model ensures that the LIFE-SHARE Project identifies costs and institutional/consortial strategies for all aspects of digital content curation and preservation.

Across the digital content lifecycle, required skills range from project management to the preparation of exacting technical specifications, and from assessing the needs of users to constructing media-specific metadata profiles. To add to this complexity, different skills are drawn upon within different contexts – digitisation of printed material within the day-to-day activities of an academic library is likely to be a very

different to a special-funded and time-limited digitisation project. The status, nature and condition of materials may also call for highly bespoke digitisation skills. The 2007 JISC Digitisation Conference at Cardiff [2] emphasised the need to describe the nature of these interdisciplinary skills along with current training provision. The LIFE-SHARE Project directly addresses this need by working with project partner, JISC Digital Media, to assess the skills required for the creation, curation and preservation of digital content and match these requirements with current training provision available to the UK's academic community.

The Ithaka report, 'Sustainability and Revenue Models for Online Academic Resources' [3] identified compelling reasons to collaborate on digital content creation, curation and management – both within and across institutions. The LIFE-SHARE Project, being a consortial project, is well-placed to explore both institutional and consortial strategies to support digitisation activities, and the wealth of expertise across the partner institutions will ensure that the project outcomes will capitalise on the pooling of experiences – for the benefit of the wider community.

2 Methodology

The design of the Digitisation Strategy Toolkit involves an iterative process. The first stage is a digitisation audit; this is followed by the drafting of a provisional toolkit. Case studies at the three institutions form the second stage of the project, focusing on different aspects of the digital lifecycle. The work from these case studies is then used to inform the next version of the toolkit. In the third stage, the project addresses the question of consortial models for offering digitisation services. The work from this will contribute to the creation of the final version of the toolkit. Throughout this process, the toolkit has been circulated to digitisation practitioners in the Higher Education community for evaluation and comment.

3 Digitisation Audit and Skills Map

An audit of all digitisation across the three White Rose Libraries has been completed. The audit took the form of a review of digitisation services, and an inventory of all activities involving digitisation across the lifecycle. The review of services revealed that the three universities offer very similar services, particularly on-demand digitisation for access to course reading materials and for access to archive materials and project based digitisation for selected collections. However, the development of the services differed significantly between the institutions.

The inventory of all digitisation activities across the three institutions provides an invaluable resource. It contains lists of staff members' skills and expertise, equipment, strategies, digital collections and potential collections. The inventory supports the results of the services review in that there are significant amounts of similar work occurring across all three institutions. It has also revealed some gaps in the provision of service and expertise, such as the ability to scan materials larger than A2 and a lack of capability in the area of Encoded Archival Description (EAD).

The skills and training maps have been developed to enable users to identify the skills needed for the type of digitisation they are planning. Users can carry out a knowledge check to help inform their choice of training.

4 Case Studies

There are three case studies, one based at each of the three White Rose institutions. For each study there is a particular emphasis on a slightly different stage of the digital lifecycle, though all take into consideration the issues involved at every stage.

The Leeds case study investigates the costing and workflow for digitising print monographs for preservation purposes. The focus is on a comparison between the costs of the conservation of physically deteriorating monographs from the early twentieth century and the costs of producing and preserving digital copies of the same monographs.

The Sheffield case study investigates the workflow for digitising audio and video recordings from Special Collections with a particular focus on the permissions required for preservation and dissemination.

The York case study investigates the workflow for providing on-demand digitisation services for online course readings and archive materials. The study explores the two strands of on-demand digitisation and develops workflows for each strand. Different approaches to the ways in which the two strands may collaborate and combine services have also been considered.

5 Consortial Models

This stage of the project draws on the first 2 stages to establish whether the White Rose University Libraries could use shared expertise and equipment to provide a consortial digitisation service. A number of different models have been created to address both on-demand and project based digitisation. The models allow the institutions to address any gap in service or expertise by drawing on the consortium.

6 Strategy Toolkit

The final stage of the project draws on all previous stages to create the Digitisation Strategy Toolkit. This work has been carried out concurrently with the other stages to enable the development and updating of the toolkit as the project progresses.

The toolkit essentially provides a framework for producing a digitisation strategy. This uses the structure of the LIFE model, exploring each stage of the digital lifecycle: creation, acquisition, ingest, content preservation, bit-stream preservation and access. For each of the lifecycle elements the toolkit looks at the policies, infrastructure, and resources (internal and external) necessary to support digitisation strategy.

6.1 Policies

This section outlines the existing Library policies that can inform the creation of a digitisation strategy. These policies may provide source material for the strategy, or

the digitisation strategy may be written into these policies. Where possible the toolkit links to examples of existing university policies available online.

6.2 Infrastructure

This section outlines the infrastructure necessary to support digitisation. Most of the items under this section will be unique to the institution, such as staff or available equipment. However, where it is possible, sources of infrastructure that are available to multiple institutions, for example JORUM, are linked to.

6.3 Resources

This section draws most heavily on the work of the case studies, audit and skills maps. The resources offer reusable best practice guidelines, technical standards and costings for undertaking digitisation, created from our experiences in the case studies. The equipment and staff lists from the audit also form a core part of the internal resources. Whilst these are lists are unique to the White Rose institutions, the methodology for creating them can be replicated at other institutions. The skills map provides a workflow for anyone undertaking a digitisation project, assessing the user's skills and providing links to a wide range of external training materials.

7 Conclusions

The creation of a digitisation strategy is an essential part of the work libraries must carry out in order to manage their digital and print collections in the long term. The strategy should be based on a thorough understanding of both existing digitisation work and potential future digital collections. It is important to integrate the digitisation strategy with other existing policies such as those relating to preservation and copyright and IPR. The strategy must also address all stages of the digital lifecycle, from creation through ingest to long term preservation. The LIFE-SHARE Digitisation Strategy Toolkit provides a useful model for institutions to use, that will enable the development of a fully integrated policy as well as providing links to external sources of best practice, advice and training.

References

1. Wheatley, P., Ayris, P., Davies, R., Mcleod, R., Shenton, H.: The LIFE Model v1.1. LIFE Project, London, UK (2007), <http://eprints.ucl.ac.uk/4831/>
2. JISC Digitisation Conference (2007), http://www.jisc.ac.uk/media/documents/publications/digi_conference_report-v1-final.pdf
3. Guthrie, K., Griffiths, R., Maron, N.: Sustainability and Revenue Models for Online Academic Resources. An Ithaka Report (2008), http://sca.jiscinvolve.org/files/2008/06/sca_ithaka_sustainability_report-final.pdf

Ensemble: A Distributed Portal for the Distributed Community of Computing Education

Frank M. Shipman¹, Lillian Cassel², Edward Fox⁶, Richard Furuta¹,
Lois Delcambre³, Peter Brusilovsky⁴, B. Stephen Carpenter II¹,
Gregory Hislop⁵, Stephen Edwards⁶, and Daniel D. Garcia⁷

¹ Texas A&M University, College Station, TX. 77843

² Villanova University, 800 Lancaster Avenue, Villanova, PA. 19085

³ Portland State University, 2121 SW Fourth Ave., Portland, OR. 97207

⁴ University of Pittsburgh, University Club, Pittsburgh, PA. 152132303

⁵ Drexel University, 3201 Arch Street, Philadelphia, PA. 191042737

⁶ Virginia Tech, Dept. CS, M/C 0106, Blacksburg, VA. 24061

⁷ University of California, Berkeley, CA 94720-4600

cassel@acm.org

Abstract. NSF's NSDL is composed of domain-oriented pathways. Ensemble is the pathway for computing and supports the full range of computing education communities, providing a base for the development of programs that blend computing with other STEM areas (e.g., X-informatics and Computing + X), and producing digital library innovations that can be propagated to other NSDL pathways. Computing is a distributed community, including computer science, computer engineering, software engineering, information science, information systems, and information technology. Ensemble aims to provide much needed support for the many distinct yet overlapping educational programs in computing and their associated communities. To do this, Ensemble takes the form of a distributed portal providing access to the broad range of existing educational resources while preserving the collections and their associated curatorial processes. Ensemble encourages contribution, use, reuse, review, and evaluation of educational materials at multiple levels of granularity.

Keywords: Ensemble, distributed portal, distributed community.

1 Introduction: Many Disciplines and Many Communities

The computing disciplines are varied in their perspectives on problems but overlap in many of the concepts taught. For example, the following brief synopses of several disciplines, adapted from Computing Curricula 2005 Overview Report [1]: **Computer Engineering** is concerned with design and construction of computers and computer-based systems involving the study of hardware, software, communications, and the interaction among them. **Computer Science** includes design and implementation of challenging software, new uses for computers, and effective ways to solve computing problems, with emphasis on correctness and performance. **Information Systems** integrates information technology solutions and business processes to meet

the information needs of business, enabling them to achieve their objectives in an effective, efficient way. **Information Technology** trains specialists with the right mix of knowledge and hands-on expertise to take care of both an organization's information technology infrastructure and the people who use it. **Software Engineering** addresses developing and maintaining software systems that behave reliably and efficiently, are affordable, and satisfy all the requirements that customers have defined for them.

Other related disciplines include **Information Science**, which investigates information creation, processing, and use; and various kinds of informatics (e.g., **Medical Informatics**, **Bioinformatics**, etc.), where the focus is on applying computing to support the needs of (and on generating solutions for the unique demands of) specialized application areas.

The growing number of distinct computing fields and associated curricula offers the kind of richness that we need to meet the varied demands on our workforce. Each field seeks distinct goals, covers overlapping topics at varying depths, and embodies a certain perspective. In this project, we celebrate these disciplines and we leverage the years of effort devoted to articulating curricular guidelines, accreditation standards and procedures, and a common ontology [2], to identify, at the level of topics and subtopics within the various bodies of knowledge, cross-disciplinary connections.

To support this diversity of perspectives while allowing the sharing of educational content, we have developed a distributed portal that provides alternative entry points to a shared set of collections, tools, and communities. The following sections provide an overview of the distributed portal and the resources to which it provides access. We conclude with our vision for the continued evolution of Ensemble.

2 Distributed Portal

People interact with a variety of on-line communication channels for their information and Ensemble provides multiple portals (i.e., starting points for access) so that patrons can explore resource locations from their favored medium. Initially, Ensemble provides access via standard Web pages, through Facebook, and through virtual worlds (e.g., Second Life). These cover the dominant communication forms from the beginning of the Web (e.g., hypertext) and today (e.g., social networks), as well as a potential communication form of the future (e.g., virtual worlds) [3]. Each of these portals provides access to all the resources in Ensemble but presents an alternate interface to these resources. Also, some resources, such as virtual objects in Second Life, or groups or applications in Facebook, are represented through metadata in their non-native portals. Figure 1 (left) shows the Ensemble home page providing archways to collections, communities, and tools. Figure 1 (right) shows the view within the Ensemble structure in Second Life. The archway on the stand/pillar on the right is similar in functionality to the archway in the Web portal while in the middle of the view are some of the Second Life specific resources for computing education.

Ensemble can be found on the web at www.computingportal.org and in Second Life at slurl.com/secondlife/Educators%20Coop%204/73/239/28.

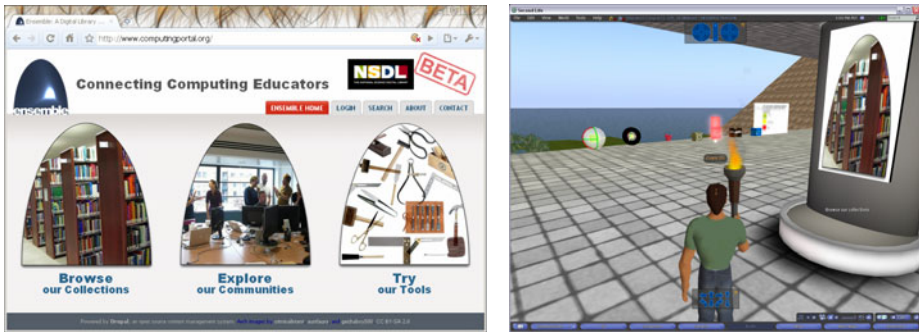


Fig. 1. The Web portal (left) and a view from within the Second Life portal (right)

3 Ensemble Resources: Collections, Tools, Communities

The distributed portal of Ensemble provides access to a range of static and evolving content in the form of existing collections, tools for education and resource development, and communities of practice.

3.1 Collections

As with many education-oriented libraries, Ensemble includes collections of traditional resources (e.g., syllabi, reading lists). Ensemble also includes a collection of tools that are valuable for education. In order to provide domain-specific access to these resources, an ontology crossing the sub-fields has been developed to index the resources. Collections currently available through Ensemble include the existing collections of AlgoViz (Algorithm Visualization), CITIDEL computing education resources [4], Computer Science Teachers Association teaching and learning materials, and Computer Science Syllabi, as well as providing a space for new collections including TECH, a collection of tools for use in education [5].

3.2 Tools and Services

Existing resources and tools only cover some of the patron's needs. Ensemble includes a number of tools that have been developed or expanded to support the creation and maintenance of digital resources. Tools like Walden's Paths [6] and the Visual Knowledge Builder [7] support the creation of meta-documents (i.e., resources made up of other existing resources), while sub-document services are being developed to enable the selection of the pieces of resources that fit a particular need. Additional Ensemble services are aimed at personalizing the presentation of resources and motivating patrons to explore and participate in Ensemble.

3.3 Communities

Communities are a resource where the voices of the members provide a continuously evolving set of content. In addition to evolving content, communities are reactive in

that, as questions or topics are brought up, answers and discussions are hopefully generated. Ensemble currently includes communities on CS1 (first programming class), the Future of Computing Education Summit (FOCES), Music and Computing, and Living In the KnowEdge Society (LIKES).

Ensemble is exploring motivating participation by awarding badges to people who contribute resources, commentary, or are otherwise active. Additionally, we are developing personalization and social navigation [8].

4 The Future Evolution of Ensemble

As described, Ensemble views a library as a source of existing content, tools to create and maintain new content, and a social structure for discussing content. The variety of portals provides patrons with a means to interact with Ensemble within their preferred medium. As we gain experience with the use of the portals and tools and as the collections and communities continue to grow, we expect the focus of Ensemble to evolve. The lessons learned from this distributed portal to a distributed community is likely to lead to new portals and tools, and to provide a model for other digital libraries crossing media and domains.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant Numbers 0534762, DUE-0840713, 0840719, 0840721, 0840668, 0840597, 0840715, 0511050, 0836940 and 0937863.

References

1. Curriculum 2005: The Overview Report, in Secondary Curriculum 2005: The Overview Report, Secondary. ACM Press (2005)
2. Cassel, L.N., et al.: The computing ontology project: the computing education application. In: Proc. of SIGCSE Symposium on Computer Science Education (2007)
3. Fox, E., et al.: Ensemble PDP-8: Eight principles for distributed portals. In: Proceedings of ACM and IEEE Joint Conference on Digital Libraries (JCDL) (2010)
4. Impagliazzo, J., Cassel, L.N., Knox, D.: Using CITIDEL as a Portal for CS Education. *Journal of Computing in Small Colleges* (2002)
5. Garcia, D., Bailes, D., Fincher, S.: Technology that Educators of Computing Hail (TECH) (Birds of a Feather). In: SIGCSE 2009, Chattanooga, TN, March 4-7 (2009)
6. Shipman, F., et al.: Using Paths in the Classroom: Experiences and Adaptations. In: Proceedings of ACM Hypertext 1998, pp. 267–276 (1998)
7. Shipman, F., et al.: The Visual Knowledge Builder: A Second Generation Spatial Hypertext. In: Proceedings of the ACM Conference on Hypertext, pp. 113–122 (2001)
8. Brusilovsky, P., et al.: Comprehensive personalized information access in an educational digital library. In: Proc. of Joint Conference on Digital Libraries (2005)

A New Focus on End Users: Eye-Tracking Analysis for Digital Libraries

Jonathan Sykes¹, Milena Dobрева², Duncan Birrell², Emma McCulloch²,
Ian Ruthven³, Yurdagül Ünal⁴, and Pierluigi Feliciati⁵

¹ Glasgow Caledonian University, Glasgow, UK
jon.sykes@gcal.ac.uk

² Centre for Digital Library Research (CDLR), University of Strathclyde, Glasgow, UK
{milena.dobрева, e.mcculloch, duncan.birrell}@strath.ac.uk

³ Computer and Information Sciences, University of Strathclyde, Glasgow, UK
Ian.Ruthven@cis.strath.ac.uk

⁴ Department of Information Management, Hacettepe University, Ankara, Turkey
yurdagul@hacettepe.edu.tr

⁵ Department of Cultural Heritage, University of Macerata, Fermo, Italy
pierluigi.feliciati@unimc.it

Abstract. Eye-tracking data was gathered as part of a user and functional evaluation of the Europeana v1.0 prototype, to determine which areas of the interface screen are most heavily used and which areas attract users' attention but are not effectively used in search. Outputs from eye-tracking data can offer insight into how advanced search functions can be made more intuitive for end users with differing interests and abilities, and can be used to inform continued interface development as digital libraries look to the future. Results led to recommendations for the future development of the Europeana digital library.

Keywords: digital libraries, eye-tracking, gaze plots, heat maps, user studies.

1 Introduction

With the focus of interface design firmly on the end user [1], research methods such as eye-tracking can be used to provide insight to their search behaviour, informing the development of effective services for users of different ages, abilities, interests and backgrounds. This paper explains how eye-tracking can add insight into the use of digital libraries not possible by other means, demonstrated here in the context of a commissioned evaluation of the Europeana¹ Digital Library, conducted in October 2009-January 2010 and coordinated by the Centre for Digital Library Research (CDLR) at the University of Strathclyde.

Europeana is a single access point for digitised cultural heritage materials provided by various European memory institutions, launched by the President of the European Commission (EC) in November 2008. It aims to have 10 million objects by the end of 2010. The Europeana v1.0 interface is available in 26 European languages; the digital

¹ <http://www.europeana.eu/portal/>

library supports both simple and advanced searching, a series of tabs and filters to refine queries, and additional functionality such as a timeline and date clouds.

The present study combined a series of traditional focus groups (77 participants in four European cities) with 12 one-to-one media labs, in which eye-tracking data was collected [2]. This paper focuses on the outcomes from the media labs and more specifically on the eye-tracking component and its value in such studies.

2 The Use of Eye-Tracking in User-Centric Studies

Eye-tracking technology enables researchers to examine how users navigate search results, and which aspects of an interface they deem most important [3]. Eye-tracking has been used extensively in psychology research, investigations of game playability [4] and has also been incorporated into user-centric studies on human information behaviour [5], studying how users, e.g., use textual and pictorial representations of video objects [6], search online [7], evaluate results lists produced by search engines [8], evaluate different results screen interfaces (tabular and list) [9], make decisions relating to Google results following a query [10], and how task and gender influence search and evaluation behaviour [11]. Although this method could provide useful insights about the use of electronic resources, it is still not popular in user-oriented evaluations within the digital library domain.

3 Methodology and Findings

The study ensured a uniform methodology across focus groups and media labs. Collected data included three questionnaires (demographic data, initial and lasting impressions of Europeana), discussions validating questionnaire-based feedback, and populated PowerPoint presentations combining a set of scenarios which required users to formulate searches that targeted a range of Europeana's metadata fields.

The eye-tracking studies were conducted to scope 'areas of interest' (AoI) and sequences of actions of participants in the study. Data gathered were visualised as heat maps and gaze plots of test participants' eye movements as they interacted with the Europeana interface and analysis of AoIs for three types of interface screens of Europeana (home screen, search results and timeline); these are presented in detail in the final report of the study [2]. The study made 24 recommendations which related to the content, functionality/usability and navigation. The analysis of the eye tracking data was essential for the recommendations related to navigation and was useful for some of the recommendations related to functionality/usability. Thus this method provided additional insight into the way participants were using Europeana, which complemented the rest of the data gathered within the study.

Here we present only one example of the user tracking data and their analysis. Fig 1 shows that, currently, the images presented on the home screen suffer from issues of low saliency. They are rarely noted during initial exposure, accounting for just 4% of user fixation. If the images are to serve a design purpose, it seems they are currently failing.



Fig. 1. a. Europeana 'home screen' augmented by AoI. **b.** Doughnut graph showing percentage of fixation for each AoI on 'home screen'.

The study of the search screen showed that the weighting between the three results features was comparable – images 23% of fixation; image navigation 15%; refined search 15%. This suggests the screen layout is suitably balanced, excepting the limited salience of the top navigation bar.

The use of the timeline screen which allows users to browse images using a single clickable navigation point had also been analysed. Few participants made full use of it; some participants stated that they did not see the date clouds, representing the number of tagged objects for a given date, as this display required users to scroll down in the browser.

4 Conclusions

This study's methodology enabled feedback from participants to be compared with tangible data captured on their actual information seeking behavior and the search strategies deployed throughout the assignment. Most had not consciously considered what search options they had looked at, used or rejected on the interface during a single search session; additionally, they had not realized their frequent repetition of similar (often unsuccessful) search strategies or terms deployed across the range of tasks/scenarios, thus validating the value of eye-tracking methodology. Data indicated that the home screen could be better balanced, with more prominence given to images; the navigation bar on the result screen could be more arresting and the timeline feature would be better arranged on a single screen, with no need for scrolling.

Some problems and limitations with the use of eye-tracking technology exist and should be considered: lack of in-house software may necessitate outsourcing; the method is time consuming and more intrusive than more traditional methodologies, thus requiring ethics approval; and the capture technology cannot (at present), be universally applied to all end users, as those with eye defects cannot make suitable subjects for research.

Limitations aside, the use of eye-tracking data offers evidential insight into the psychological perception of the parameters of advanced search and the physiological tracking of this encounter – where end users were recorded as they engaged and

disengaged (often in rapid succession) with a range of search options available to them on the interface. It provides an innovative means of encouraging reflective practice in end users and the development community; and it can help to plot the co-ordinates of continued development as digital libraries look to the future.

References

1. Kani-Zabihi, E., Ghinea, G., Chen, S.Y.: Digital Libraries: what do users want? *Online Information Review* 30, 395–412 (2006)
2. Dobрева, M., McCulloch, E., Birrell, D., Feliciati, P., Ruthven, I., Sykes, J., Ünal, Y.: User and Functional Testing. Final report. Europeana v. 1.0. Final report. 63 pp. Appendices. 104 pp. (2010), <http://version1.europeana.eu/web/europeana-project/documents>
3. Granka, L., Feusner, M., Lorigo, L.: Eye Monitoring in Online Search. In: Hammoud, R. (ed.) *Passive Eye Monitoring*, pp. 283–304. Springer, Heidelberg (2008), http://laura.granka.com/publications/granka_etal08book.pdf
4. http://www.lutin-userlab.fr/_pages/english/
5. Klami, A., Saunders, C., Campos, T., Kaski, S.: Can relevance of images be inferred from eye movements? In: *MIR 2008: Proc. 1st ACM Int. Conf. on Multimedia Information Retrieval*, Vancouver, British Columbia, Canada, October 30-31 (2008)
6. Hughes, A., Wilkens, T., Wildemuth, B.M., Marchionini, G.: Text or Pictures? An Eyetracking Study of How People View Digital Video Surrogates. In: Bakker, E.M., Lew, M., Huang, T.S., Sebe, N., Zhou, X.S. (eds.) *CIVR 2003. LNCS*, vol. 2728, pp. 271–280. Springer, Heidelberg (2003)
7. Granka, L.A.: *Eye-R: Eye-tracking Analysis of User Behavior in Online Search*. Masters Thesis, Cornell University Library Press (2004), http://laura.granka.com/publications/granka_MSthesis04.pdf
8. Aula, A., Majoranta, P., Rähkä, K.-J.: Eye-tracking Reveals the Personal Styles for Search Result Evaluation. In: Costabile, M.F., Paternò, F. (eds.) *INTERACT 2005. LNCS*, vol. 3585, pp. 1058–1061. Springer, Heidelberg (2005), <http://www.cs.uta.fi/~curly/publications/INTERACT2005-Aula.pdf>
9. Rele, R.S., Duchowski, A.T.: Using Eye-tracking to Evaluate Alternative Search Results Interfaces. In: *Proceedings of the Human Factors and Ergonomics Society*, Orlando, Florida, pp. 1459–1463 (2005)
10. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, article 3 12(3) (2007), <http://jcmc.indiana.edu/vol12/issue3/pan.html>
11. Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., Gay, G.: The influence of task and gender on search and evaluation behavior using Google. *Information Processing & Management* 42, 1123–1131 (2006), http://laura.granka.com/publications/lorigo_etal.pdf

Digital Library Educational Module Development Strategies and Sustainable Enhancement by the Community

Seungwon Yang, Tarek Kanan, and Edward Fox

Department of Computer Science, Virginia Tech
Blacksburg, VA 24061 U.S.A.
{seungwon, tarekk, fox}@vt.edu

Abstract. The Digital Library Curriculum Development Project (<http://curric.dlib.vt.edu>) team has been developing educational modules and conducting field-tests internationally since January 2006. There had been three approaches for module development in the past. The first approach was that the project team members created draft modules (total of 9) and then those modules were reviewed by the experts in the field as well as by other members of the team. The second approach was that graduate student teams developed modules under the supervision of an instructor and the project team. Four members in each team collaborated for a single module. In total four modules were produced in this way. The last approach was that five graduate students developed a total of five modules, each module reviewed by two students. The completed modules were posted in Wikiversity.org for wider distribution and collaborative improvements by the community¹. The entire list of modules in the Digital Library Educational Framework also can be found in that location.

Keywords: digital libraries, curriculum, education, module development, development strategy, wiki.

1 Introduction

The Digital Library (DL) Curriculum Development Project² created a united curriculum that can be used for both CS and LIS disciplines [1, 2, 3]. In this four-year joint project³ between the Dept. of CS at Virginia Tech and the School of Information and Library Sciences at University of North Carolina at Chapel Hill, a total of 19 educational modules out of 47 in the DL Module Framework have been developed based on three different strategies. In this poster, we present developed modules, strategies used, and the module developers' comments on their experience with the third strategy. The effort to develop and evaluate modules is ongoing. The completed modules have been posted to Wikiversity.org¹ and related Wikipedia articles are linked to the

¹ http://en.wikiversity.org/wiki/Curriculum_on_Digital_Libraries

² <http://curric.dlib.vt.edu>

³ This collaborative project was supported by the National Science Foundation under Grant Nos. IIS-0535057 (VT) and IIS-0535060 (UNC-CH).

modules for sustainable enhancement by the interested members of the community and for wider use by the public.

2 Developed Modules from DL Curriculum Framework

Figure 1 presents the developed modules with three strategies in different colors. They show a broad coverage of topics in the DL area (9 core topics out of 10).

FRAMEWORK FOR A DIGITAL LIBRARY CURRICULUM		
CORE TOPICS		Colors: strategy I , strategy II , strategy III
1	Overview	1-a (10-c): Conceptual frameworks, models, theories, definitions 1-b: History of digital libraries and library automation
2	Digital Objects	2-a: Text resources 2-b: Multimedia 2-b(1): Images 2-c (8-c): File formats, transformation, migration
3	Collection Development	3-a: Collection development/ selection policies 3-b: Digitization 3-c: Harvesting 3-d: Document and e-publishing/ presentation markup 3-e (7-e): Web Publishing 3-f (7-f) Crawling
4	Info/ Knowledge Organization	4-a: Information architecture (e.g., hypertext, hypermedia) 4-b: Metadata 4-c: Ontologies, classification, categorization 4-d: Subject description, vocabulary control, thesauri, terminologies 4-e: Object description and organization for a specific domain
5	Architecture (agents, mediators)	5-a: Architecture overviews 5-b: Application software 5-c: Identifiers, handles, DOI, PURL 5-d: Protocols 5-e: Interoperability 5-f: Security
6	User Behavior/ Interactions	6-a: Info needs, relevance 6-b: Online information seeking behavior and search strategy 6-c: Sharing, networking, interchange (e.g., social) 6-d: Interaction design, usability assessment 6-e: Info summarization and visualization
7	Services	7-a: Search engines, IR, indexing methods 7-a (1): Image retrieval 7-b: Reference services 7-c: Recommender systems 7-d: Routing, community filtering 7-e (3-e): Web publishing 7-f (3-f): Crawling 7-g: Personalization
8	Preservation	8-a: Preservation 8-b: Web archiving 8-b: Sustainability 8-c (2-c): File formats, transformation, migration
9	Management and Evaluation	9-a: Project management 9-b: DL case studies 9-c: DL evaluation, user studies 9-d: Bibliometrics, Webometrics 9-e: Intellectual property 9-f: Cost/economic issues 9-g: Social issues
10	DL education and research	10-a: Future of DLs 10-b: Education for digital librarians 10-c (1-a): Conceptual framework, theories, definitions 10-d: DL research initiatives

Fig. 1. The DL Module Framework with developed modules in color

3 Development Strategies

There had been three approaches for module development in the past. The first approach was that the project team members created draft modules (Strategy I in Table 1) and then those modules were reviewed by the experts in the field as well as by other members of the team. The second approach was that graduate student teams developed modules under the supervision of an instructor and the project team. Four members in each team collaborated for a single module (Strategy II in Table 1). Those modules were taught in a graduate level digital library course.

Table 1. Three strategies and corresponding modules developed

Strategies	Modules	Developed by	Reviewed by	Taught by
Strategy I	1b*, 3b, 4b, 5a, 5b, 6a, 6b, 6d, 7b, 9c	Project team	Field experts	Class teams/ instructor (*)
Strategy II	2c, 5d, 7g, 8b	Class teams	The other teams	Class teams
Strategy III	7a, 7c, 7d, 7f (3f), 8a	Grad students	Team peers	Grad students

As our third approach, five graduate students in the Dept. of CS at Virginia Tech developed (IR-related) modules throughout the CS 5604 Information Retrieval class in Fall 2009 (Strategy III in Table 1). A description of their experience was collected from a survey and is presented in Table 2; it will be used to help future developers.

Table 2. Developers' comments on their experience from Strategy III

Categories	Comments
Most challenging sections	<ul style="list-style-type: none"> • 5S Characteristics (defines the streams, structures, spaces, scenarios, and societies aspects of the module) <ul style="list-style-type: none"> ○ Difficult to categorize the topic into 5 characteristics. ○ Difficult to understand 5S to fit them into a module. • Exercises and Learning activities. <ul style="list-style-type: none"> ○ It is hard if you don't have suitable knowledge about the topic. ○ It's not easy to create good, entertaining, comprehensive and active exercises.
Teaching the modules in class	<ul style="list-style-type: none"> • The module is a good starting point. • Start from reading the body of knowledge section then focus on specific parts based on the resources section.
Suggestions for future developers	<ul style="list-style-type: none"> • Find good resources and take a general look on them. • Start from the learning objectives then do the body of knowledge. • Develop the evaluation of learning objectives. • Incorporate informational and entertaining exercises and learning activities.

Under the supervision of the project team and the course instructor, who is also on the project team, each student selected a module to work on and then began completing the sections individually. The complete list of sections in a module template⁴ can

⁴ <http://curric.dlib.vt.edu/modDev/Template.2008-10-03.pdf>

be found at the project homepage. During the first half of the semester, they focused on the module name, its scope, learning objectives, taxonomic characteristics, relationships with other modules, prerequisite knowledge required, a list of topics for the body of knowledge, resources, and additional useful links. This was to produce a basic overall structure of the modules. Comments were provided by the project team following the students' midterm presentation of their draft modules. During the final half of the semester, students completed sections covering the level of effort required, details of the body of knowledge, exercises and learning activities, evaluation of learning objective achievements, and glossary terms. Each module was reviewed by two other students in a non-overlapping way.

4 Sustainable Enhancement on Modules

The completed modules have been posted to Wikiversity.org to ensure continuous improvement by interested members of the community. Related Wikipedia articles are linked to the modules to provide further details. We plan to invite experts in the field so that they could review the modules and leave comments. As shown in our previous module evaluation, this use of a wiki for evaluation will allow asynchronous communication among the evaluators. They can read other evaluators' comments for the same module and express their (dis)agreement on the wiki discussion page, as well as add their own comments.

5 Conclusion

We can enrich module development outcomes by using different development strategies and posting the modules to Wikiversity.org, ensuring continuous enhancement by interested members of the digital libraries community. Based on the graduate students' comments, we note that there are some module sections that were more difficult to develop than other sections; we can provide help. On the other hand, teaching using the developed modules can be helpful and interesting for instructors.

References

1. Pomerantz, J., Wildemuth, B., Fox, E.A., Yang, S.: Curriculum Development for Digital Libraries. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 175–184. ACM, New York (2006)
2. Yang, S., Fox, E.A., Wildemuth, B., Pomerantz, J., Oh, S.: Interdisciplinary Curriculum Development for Digital Library Education. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 61–70. Springer, Heidelberg (2006)
3. Yang, S., Wildemuth, B., Kim, S., Murthy, U., Pomerantz, J., Oh, S., Fox, E.A.: Further Development of a Digital Library Curriculum: Evaluation Approaches and New Tools. In: The 10th International Conference on Asian Digital Libraries, Hanoi, Vietnam, December 10-13 (2007)

Approach to Cross-Language Retrieval for Japanese Traditional Fine Art: Ukiyo-e Database

Biligsaikhan Batjargal¹, Fuminori Kimura², and Akira Maeda²

¹ Graduate School of Science and Engineering, Ritsumeikan University, Japan
biligsaikhan@gmail.com

² College of Information Science and Engineering, Ritsumeikan University, Japan
{fkimura, amaeda}@is.ritsumei.ac.jp

Abstract. In this paper we introduce our system that retrieves Ukiyo-e databases using an English query by customizing and utilizing freely available open source software. In our system, the Ukiyo-e metadata elements were mapped to Dublin Core. We adopted a dictionary-based query translation approach and utilized the Greenstone Digital Library Software to make available our Ukiyo-e digital collections online. The preliminary result is an easy-to-use and useful system for users who do not understand Japanese, that allows to search and view Japanese Ukiyo-e databases in English.

Keywords: Ukiyo-e, Image database, Digital library, Cross-Language information retrieval.

1 Introduction

Our goal is to develop an easy-to-use system for users who do not understand Japanese, which allows to search and browse Japanese Ukiyo-e databases in English through a single interface and a single query.

The Ukiyo-e, Japanese traditional woodblock printing is known world-wide as one of the fine arts in the Edo period (1603–1868) which originated in the metropolitan culture of Tokyo during Edo. In recent years, the role and importance of digital cultural heritage preservation has been continuously increasing. Many Ukiyo-e prints are being digitized and made publicly available.

The big collections of Ukiyo-e such as Museum of Fine Arts of Boston and Tokyo National Museum offer searching and browsing features by few metadata fields. However, the needs of humanities scholars and researchers are diverse that might require accessing more detailed information rather than searching and browsing few collections in one language. Unlike most of the Ukiyo-e collections, few image databases including the Ukiyo-e collections of the Victoria and Albert Museum collections and the Ukiyo-e database of the Art Research Center of Ritsumeikan University offer extensive search features with additional metadata fields in their rich data. The collections in Japanese institutions are available only in Japanese, so that users who do not understand Japanese may not find the desired information. Besides, the texts of Ukiyo-e databases such as special terms, names of artists, etc. contain

archaic Japanese words, which reflect the Japanese language of the Edo period. We have started to develop a system, which retrieves Japanese Ukiyo-e databases using an English query by utilizing the achievements of cross-language information retrieval (CLIR).

2 Related Work

Much research has been conducted in the past on CLIR. However, little research has tried to apply achievements of CLIR to ancient languages. Recently, some research conducted regarding the information retrieval of historical documents in ancient languages. For instance, Khaltarkhuu et al. [1] proposed a retrieval technique that considers cross-period differences in dialect, spelling and writing systems of modern and traditional Mongolian. The retrieval method of Kimura et al. [2] considered the language differences between ancient and modern Japanese. Koolen et al. [3] considered the spelling and pronunciation differences between ancient and modern Dutch, while Pilz et al [4] considered spelling variations of English and German historical texts. In general, the main challenge for historical European languages is the spelling variants. However, the situation for Japanese language is different because Kanji characters used to represent words and a modern character is not the same as the archaic one. An archaic word that consists of a single Kanji might be formed as more than a single modern Kanji or vice a versa. Any Kanji character has many pronunciations and used differently for words that consist of two or more Kanji characters. Besides, Japanese documents are written without explicit word boundaries. It is a rather challenging task to find information using a modern language query from Japanese documents that are written in texts mixed with modern and archaic Japanese words.

3 Proposed System

The system architecture of the proposed system is illustrated in Fig. 1. We utilized the Ukiyo-e image database of the Art Research Center of Ritsumeikan University, which is freely accessible in Japanese. At the first stage, 67 metadata elements of the Ukiyo-e database were mapped to the 11 Dublin Core metadata elements with 4 qualifiers. The rest of the Ukiyo-e metadata elements were added as the additional elements without omitting, since these are valuable for researchers. Moreover, newly created elements “readingRomaji”, “readingHiragana”, “JapaneseDate”, etc. together with the additional elements were used for browsing and displaying Japanese content in English. KAKASI – Kanji Kana Simple Inverter was utilized to convert Kanji characters to Hiragana or Latin alphabet. MeCab – Part-of-speech and morphological analyzer for Japanese was used for word-segmentation. In general, the Ukiyo-e database was processed that archaic and modern names of Ukiyo-e artists, performers, schools, etc. are segmented properly. A sample bilingual dictionary with the Ukiyo-e terms, special words and names of the artists, schools and seals was created using the Internet resources, e.g., Wikipedia. This dictionary also was utilized for the romanization, word-segmentation, GUI translation as well as the retrieval. Once the Ukiyo-e database was processed, it was imported into Greenstone as the XML files and indexed

The screenshot displays a search interface for the Ukiyo-e Database. At the top, a search bar contains the query "Kuniyoshi" and a "Begin Search" button. A search results table is visible, listing items with details such as Accession Number, Title, Performance, and Actor. A large image of a Ukiyo-e print is shown on the left side of the interface. Three callout boxes with red arrows point to specific parts of the interface:

- Box 1: Points to the search bar, stating "The English translation or pronunciation is displayed." (referring to the "Kuniyoshi" query).
- Box 2: Points to the search results table, stating "A user inputs a query in English and starts to retrieve. For instance: The Ukiyo-e artist named 'Kuniyoshi'" (referring to the search process).
- Box 3: Points to the search results table, stating "Our system translates an English query to ancient Japanese and retrieves in the Japanese Ukiyo-e database. The Ukiyo-e artist name 'Kuniyoshi' in Japanese Kanji is '国芳'." (referring to the translation of the query to Japanese characters).

Fig. 2. Searching Japanese Ukiyo-e Databases using an English query

Acknowledgments. This work was supported by the Grant-in-Aid for the Global COE Program "Digital Humanities Center for Japanese Arts and Cultures" from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References

1. Khaltarkhuu, G., Maeda, A.: Developing a Traditional Mongolian Script Digital Library. In: Buchanan, G., Masoodian, M., Cunningham, S.J. (eds.) ICADL 2008. LNCS, vol. 5362, pp. 41–50. Springer, Heidelberg (2008)
2. Kimura, F., Maeda, A.: An Approach to Information Access and Knowledge Discovery from Historical Documents. In: Digital Humanities 2009, pp. 361–363. ADHO (2009)
3. Koolen, M., Adriaans, F., Kamps, J., Rijke, M.: A Cross-Language Approach to Historic Document Retrieval. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsirikia, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 407–419. Springer, Heidelberg (2006)
4. Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P., Archer, D.: The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic. *Literary and Linguistic Computing* 23(1), 65–72 (2008)

Open Source Historical OCR: The OCRopodium Project

Michael Bryant, Tobias Blanke, Mark Hedges, and Richard Palmer

Centre for e-Research, King's College London

{michael.bryant,tobias.blanke,mark.hedges,richard.d.palmer}@kcl.ac.uk

Abstract. In this paper we present some initial results of OCRopodium project to build a scalable workflow for OCR of historical collections. Large-scale digitisation projects dealing with text-based historical material face challenges that are not well-catered-to by commercial software. Open source tools allow for better customisation to match these requirements, particularly with regard to character model training and per-project language modelling.

1 Introduction: The Problem

Large-scale digitisation projects dealing with text-based historical material face many specific challenges that are not well-catered-to by commercial software. Poor-quality and age-degraded source material, archaic languages and great variation in page-layout tend to demand a very flexible approach to OCR, and inevitably a large degree of manual intervention. Yet, we believe that there is significant benefit to be had from the development of standard workflows that increase efficiency, reduce dependence on 'black-box' commercial software, and take advantage of recent advances in open-source OCR.

This demonstration paper presents initial results of the OCRopodium project and a workflow to support scalable OCR of historical document collections. To this end, we introduce in Section 2 our general architecture and look at ways to enhance the core OCRopus software for better results with specific collections. In Section 3, we show how the results of manual correction and review can be used to iteratively improve the raw machine-generated OCR output.

2 Customising OCRopus for Historical OCR

The open-source OCRopus OCR framework [1] affords us a great deal of flexibility in constructing workflows around its core interfaces. Figure 1 summarises our architecture for scalable historical OCR, utilising a task scheduler for running conversion and training jobs in parallel, a digital repository for storing transcripts and associated metadata, and a web-based front end for administration.

A new and rapidly evolving collection of OCR interfaces, OCRopus' strength derives from its adaptability and potential for customisation. Broadly, we can provide beneficial customisations in four areas: preprocessing, page segmentation, line recognition and language modelling.

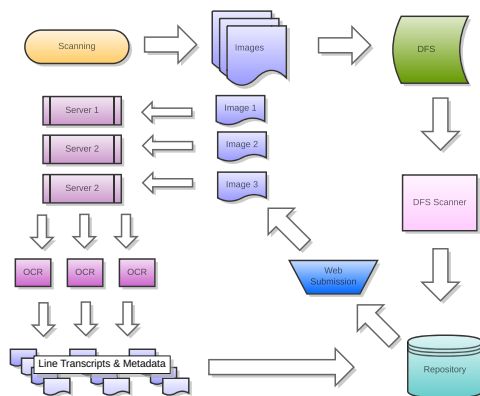


Fig. 1. OCRopodium architecture

2.1 Preprocessing

OCRopus provides a standard set of pre-processing components and an abstract interface for “plugging-in” bespoke ones, such as de-warping and de-noising. On top of this we plan to allow a more generic set of image transformation operations, and a preset system that allows institutions to define “macro” operations applicable to batches of scans exhibiting the same pre-binarisation artefacts.

2.2 Page Segmentation

OCRopus contains several different algorithms for segmenting a page into reading-order blocks of lines. Currently missing from the its tool-set, however, is a page segmenter tuned for generating good results on tabular data, a gap we aim to fill by writing a custom component. In addition, by taking advantage of OCRopus’ high-level Python APIs to integrate the segmentation components tightly into our custom workflow, we will also be able to make it more context-aware, allowing users to override certain behaviours or force segmentation with a particular page-layout scheme at a per-project level.

2.3 Line Recognition

For recognising text at an individual line-level OCRopus provides a character model that is fully trainable, albeit via a slow and computationally intensive process that requires a large amount of ground-truth training data (line images and their respective character-accurate transcripts.) As described in the next section, we believe that by preserving the positional metadata derived from the initial OCR results and combining it with the manually corrected transcript, the generation of ground-truth data for training can be made almost automatic.

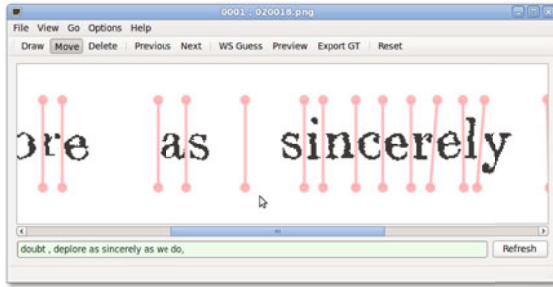


Fig. 2. Interface for line-transcript correction and character segmentation

2.4 Language Modelling

OCRopus currently provides a lexicon of possible word matches in OpenFST [\[2\]](#) format, derived from a simple dictionary list. By providing an interface to the low-level OpenFST-based tools, we'll enable institutions to more easily tailor the language modelling to their requirements by entering new words, names, and places - or potentially whole other languages - into per-project lexicons. Common-case OCR errors can be reduced by assigning positive weightings to frequently mis-transcribed words, and by preserving the results of the language modelling stage in the per-line metadata of the output transcript we can also provide better automatic error detection, making it easy to find lines containing words which failed to match the language model.

3 Building an Interface to Facilitate OCRopus Training

At present, the facilities provided by OCRopus for character model training are quite unstable and lacking a user-friendly interface. We have therefore tailored it for use in a typical historical archive workflow, with the intention of making improved character model training almost a side-effect of the normal OCR, correction, and review process:

1. A certain amount of input material is processed using the default (or best previously-determined) settings to generate per-line images and their respective machine-generated (imperfect) transcripts.
2. The output material, along with per-line positional metadata, is saved to a digital repository with an appropriate “pending” status.
3. Material is corrected manually at a per-line level (rather than per-page) since it is important to maintain the link between individual line images and their corresponding transcripts. We see this process not differing at all from that of the normal correction and review procedure.
4. The operator submits a training job to the task queue for processing.
5. The material with an approved status is retrieved from the repository. Since per-line positional metadata has been preserved, the system can link the

transcript to the area of the source image to which it refers and consider it as ground-truth text.

6. OCRopus' automated character alignment facilities are used to generate character-segmentation images, in which each letter is uniquely encoded. This is a fallible process and prone to imperfect results, due to unexpected character overlap, artefacts, and other non-text elements.
7. The operator either manually corrects the mismatched character segmentation lines, or removes them altogether from the training pool.
8. The training task proceeds, resulting in an enhanced character model.

In this manner the quality of the character model can be iteratively improved over the course of a digitisation project, reducing the word error rate (WER) and the amount of manual correction needed. Whilst early testing has shown it is difficult to better the performance of OCRopus' standard Latin character model classifier, it will significantly lower the technical barriers facing institutions that need to train OCRopus on material featuring non-Latin character sets, unusual or stylised fonts, and documents with widespread but consistently-patterned artefacts.

For experimentation with OCRopus training we have developed a standalone application, shown in Figure 2, that provides an interface for per-line transcript correction and manual character segmentation. For datasets that use a relatively standard, non-cursive font-face, we think that the burden of these tasks could be eliminated entirely.

4 Conclusion

Using the approach outlined above we believe it is possible to leverage open-source software to produce a robust, scalable system for historical digitisation projects. By automating as much as possible and providing a consolidated interface to the conversion and training components of the OCR back-end, digitisation staff will be more able to concentrate on those tasks for which they are most needed: correcting and reviewing output. Lowering the technical barriers to effective character- and language-model training will empower institutions to share experience and build on successive projects, and a rigorous emphasis on metadata preservation throughout the workflow will enable the storage of OCR outputs to be compatible with data-preservation best-practises.

References

1. Breuel, T.M.: The ocropus open source ocr system (2007)
2. The openfst project, <http://www.openfst.org>

A Voice-Oriented Image Cataloguing Environment

José H. Canós, Carlos J. Castillo, Pablo Muñoz , Héctor Valero, and Manuel Llavador

Dept. of Computer Science (DSIC), Technical University of Valencia,
Camino de Vera s/n, E46022, Valencia, Spain
jhcanos@dsic.upv.es

Abstract. VOICE is a tool for cataloguing digital images using a voice-based user interface. The goal of VOICE is to ease the introduction of descriptive metadata associated to single images or collections of pictures, so that the data entered can be used later for keyword-based image retrieval. We have developed two versions of the tool, standalone VOICE and VOICE4Picasa. The latter is an add-in to Picasa which calls the former without need to switch from one application to the other one. In our demonstration, we will show the features of both systems, adding metadata to pictures and using Picasa's retrieval features to find images in our collections.

Keywords: Image Cataloguing, Voice-based Interfaces, Speech Recognition.

1 Introduction and Motivation

Picture collections are among the most frequent multimedia content in personal computers. The versatility of modern digital cameras, as well as the low cost of storing digital pictures on hard disks, has made people hold large amounts of digital images. However, the easiness of storing picture collections contrasts with the high difficulty of retrieving specific pictures from these collections. Content-based image retrieval systems are still at an experimental stage, far from providing satisfactory performance. As an alternative, users would like to have some (descriptive-) metadata-based retrieval facilities available.

There are two sources of digital picture metadata: the digital camera and the users of the picture viewers. Digital cameras often add some technical metadata based in the EXIF standard [1] to the picture. Advanced cameras can provide also geo-referenced data, which can be used in applications like Panoramio [2]. In addition, some picture viewers offer facilities to the picture metadata editing, adding descriptive data not present in the EXIF element set. These facilities are rather basic, normally consisting of a simple form to give values to single picture metadata elements such as author, description, and the alike. Given the large number of pictures an average user holds, one can imagine how tedious an image cataloguing process can be. As a consequence, few people add descriptive metadata to their picture collections, making it very difficult to retrieve specific images in a way other than browsing.

Some additional help to image cataloguing is needed. In this demonstration, we illustrate how the use of a voice-based interface reduces the time to cataloguing by avoiding the tedious metadata typing processes. We introduce VOICE (Voice-Oriented

Image Cataloguing Environment), a simple utility which allows cataloguing pictures and collections of pictures with basic descriptive metadata using a speech-based input system. Using VOICE, a user can enter image descriptions without any keystroke, resulting in a very convenient process.

We have implemented two versions of VOICE, namely the Standalone VOICE and VOICE4Picasa. The former is a Windows application, whereas the latter has been included in Picasa [2] to take profit of its indexing capabilities, which yield to very fast picture retrieval.

2 Standalone VOICE

VOICE is a Windows-based application that uses the Microsoft Speech API (SAPI) [4]. Using a microphone, a user can direct the cataloguing process. Two types of descriptions are allowed: individual, for describing a single picture, and collective, which requires a previous selection of the pictures to be described. VOICE allows inserting values for the author, title, keywords and description. The VOICE main window includes the menu options plus the set of pictures under description. As some element values can be common for a collection (e.g. the author of the pictures), and others can be distinct, the user may at any moment select/deselect pictures as desired.

3 VOICE4Picasa

Describing images is not enough to have good retrieval facilities. Some additional functionality should be added to VOICE to allow metadata-based queries. Specifically, a catalog holding the pictures' metadata records should be implemented, as well as the catalog querying interface. Instead, we decided to use the retrieval capabilities of available picture managers, leaving VOICE as just a cataloguing tool (not a retrieval system). As an example, Picasa has powerful indexing capabilities that allow retrieving pictures very fast. As the descriptive metadata elements are embedded in each picture, having the images indexed by Picasa allows a rough keyword based retrieval. By rough we mean that we cannot specify a metadata element as search criterion (e.g. retrieving all the pictures having Suzy as author), but only those pictures in which the word "Suzy" appears at any place of its textual content. In order to be more specific, a namespace-like syntax can be used to assign values to elements (e.g. "author:Suzy" as the value of the field Author).

To avoid Picasa users to launch VOICE as a separate application, we developed a small extension to Picasa consisting of a button from which VOICE can be invoked. To catalog a collection of images from Picasa, a user must select the desired pictures and then click on the VOICE4Picasa button (the bottom rightmost button in Figure 1). Then, Picasa calls VOICE and passes the file names of the selected images as arguments of the call. Figure 1 shows (from background to foreground) the Picasa main window, the VOICE main window and the VOICE's image description form.

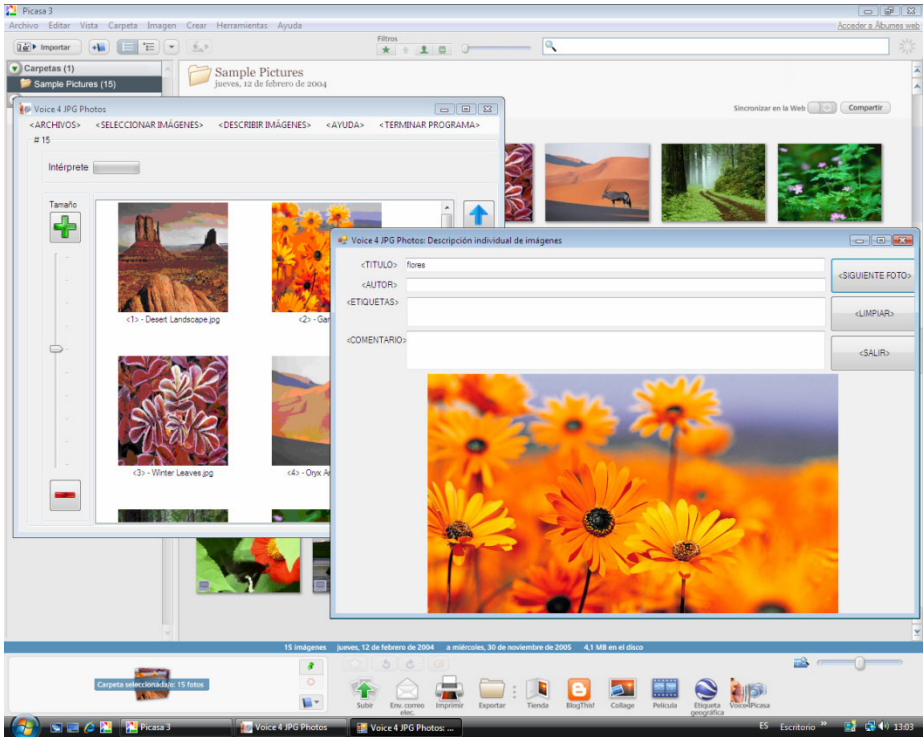


Fig. 1. Describing a picture using VOICE4Picasa

4 Outline of the Demonstration

In our Demonstration, we will illustrate the features of both VOICE and VOICE4Picasa. We will describe individual as well as sets of pictures using the speech-recognition facilities of SAPI v 3.5. Later, we will show how the catalogued images can be retrieved using Picasa. The intended audience is very wide, as no expertise is required to understand the functionality and features of VOICE.

We do not need any special equipment nor Internet connection as current version of VOICE runs in local mode.

5 Conclusions and Further Work

VOICE is an environment for voice-based metadata edition for pictures. We developed it assuming that speech-based interfaces reduce the inconveniences of typing metadata for hundreds or thousands of pictures (which is the typical size of personal image collections). We expect that features similar to those of VOICE will be added to future image managers.

Regarding VOICE limitations, we must mention that, currently, only Spanish language is supported, though the application has been designed to accept other

languages in further releases. In addition, only jpg images are supported, and other formats will be added in further releases. Finally, as the SAPI requires Windows Vista, VOICE cannot run on former versions of Windows.

Acknowledgements. The work of J. H. Canós and M. Llavador is partially funded by the Spanish *Ministerio de Educación y Ciencia* (MEC) under project META (TIN2006-15175-C05-01), and the *Junta de Comunidades de Castilla-La Mancha* under project INGENIO (PAC08-0154-9262). M. Llavador is the holder of the MEC-FPU grant no. AP2005-3356.

References

- [1] EXIF, <http://www.exif.org/>
- [2] Panoramio, <http://www.panoramio.com/>
- [3] Picasa, <http://picasa.google.com>
- [4] Microsoft Speech API 5.3,
[http://msdn.microsoft.com/en-us/library/ms723627\(vs.85\).aspx](http://msdn.microsoft.com/en-us/library/ms723627(vs.85).aspx)

DMP Online: A Demonstration of the Digital Curation Centre’s Web-Based Tool for Creating, Maintaining and Exporting Data Management Plans

Martin Donnelly¹, Sarah Jones², and John W. Pattenden-Fail²

¹ University of Edinburgh, UK

`martin.donnelly@ed.ac.uk`

² University of Glasgow, UK

`{s.jones,j.fail}@hatii.arts.gla.ac.uk`

Abstract. Funding bodies increasingly require researchers to produce Data Management Plans (DMPs). The Digital Curation Centre (DCC) has created *DMP Online*, a web-based tool which draws upon an analysis of funders’ requirements to enable researchers to create and export customisable DMPs, both at the grant application stage and during the project’s lifetime.

1 Introduction and Context

The Digital Curation Centre (DCC) defines digital curation as “maintaining, preserving and adding value to digital research data throughout its lifecycle.”^[1] The active management of research data reduces threats to their long-term research value, and mitigates the risk of digital obsolescence.

In 2009, a DCC analysis [SJ] of research funder policies and requirements for data management found that many funders “expect applicants to consider creation and management of their research outputs at the proposal stage in order to submit a data management and sharing plan.” DMP Online is a web-based tool for creating, maintaining and exporting DMPs, and has been developed in order to help research teams meet funder requirements, and respond to the recommendation in Lyon (2007) [LL] that “[e]ach funded research project should submit a structured Data Management Plan for peer-review as an integral part of the application for funding.”

The tool uses the DCC Curation Lifecycle Model [SH] as an underpinning framework to bolster its comprehensiveness; this model is designed to help researchers in defining roles and responsibilities pertaining to their data, identifying risks which arise at points of transition, and ensuring an appropriate and safe chain of custody for digital data.

¹ <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

2 Developing a Comprehensive DMP Checklist

2.1 Analysing Research Funders' Requirements and Exemplar DMPs

DMP Online is a follow-on from an earlier piece of work - the DCC Content Checklist for a Data Management Plan² - which was in turn based upon the DCC's analysis of funders' requirements and a set of exemplar DMPs.

We began by comparing what the main UK research funders ask of their applicants with regard to explicit data-related statements³, and also compared guidance produced for the UK Rural Economy and Land Use (RELU) programme⁴ and the data management guidance and manual conceived by the Australian National University (ANU)⁵.

2.2 Developing the Content Checklist for a Data Management Plan

Having analysed and synthesised the expected coverage of DMPs - and bolstered this with our own internal expertise - we suggested two iterations of such a plan; a first ('preliminary' version) for use at the grant application stage, and a second ('extended version') to be developed at the early-project stage, and updated in conjunction with the operational plan throughout the project's lifecycle.

The preliminary version (comprising those sections given in bold type in the DCC Data Management Plan Content Checklist) covers the issues that most research funders will expect researchers to address at the application stage. These typically fall into five key areas:

- What data will be created (type, format) and how;
- Plans for associated metadata and documentation, noting standards to be used;
- How data will be accessed and shared, justifying any restrictions e.g. embargoes;
- Management of Intellectual Property and ethics;
- The long-term archiving and data sharing strategy.

The extended version augments the core sections with additional information required by one or two major funders, as well as some contextual details that could usefully be included as best practice.

2.3 Public Consultation

After consulting internally among DCC colleagues, we opened the DMP Content Checklist to a public consultation via the DCC website. The clauses that

² http://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMP_checklist_v2.2_100106-publicVersion.doc

³ <http://tinyurl.com/DCC-Funder-Analysis>

⁴ <http://www.data-archive.ac.uk/relu/plan.asp>

⁵ <http://ilp.anu.edu.au/dm/>

populate DMP Online follow on from the post-consultation Checklist for a Data Management Plan (v2.2)⁶, and take into account feedback received from a variety of stakeholders via a public consultation process. The major change between the consultation document and v2.2 is that each themed paragraph has been split into a series of atomic sections, employing closed questions where possible.

3 Development of the Tool

The website and user interface were designed to enable the requirements of different funders to be mapped straightforwardly to the equivalent DCC clauses, and for onscreen guidance and links to be provided to assist in the completion of DMPs.

The tool is built atop the Ruby on Rails framework, and runs on an Ubuntu GNU/Linux server via the Apache web server. Data are stored in a MySQL database, and all technologies used in its development are free or open-source. The site is hosted by the Humanities Advanced Technology and Information Institute at the University of Glasgow, who are also responsible for the development and hosting of other digital preservation-related project sites, such as Planets, DRAMBORA, the Data Audit Framework and DigitalPreservationEurope.

Users are required to register for the site. To protect against spam-generating scripts, the tool uses the reCaptcha service to verify that users are human. From a database design perspective, ‘administrator’ users have maximum flexibility in setting up the DMP forms. Funder requirements are likely to change in time, so the system enables non-programmers to edit the mappings between individual funders and the corresponding DCC clauses. This flexibility allows for one-to-one mappings (where one funder’s requirement maps directly to one DCC wording), one-to-many mappings (where a funder’s requirement maps to multiple DCC questions), and one-to-none, for cases where there are no equivalent mappings to the DCC terms: these generally occur when the funder asks for non data-related elements to be included within a DMP (or equivalent, such as the AHRC’s Technical Appendix.)

Rather than hardcoding questions into the database, an abstract system was set up whereby questions are stored in a ‘questions’ database table. Each row of this table defines one DCC question or subject heading. The fields store the text of the question, the DCC number of the question, and a question type (text entry, true/false, or heading).

Because it is important for users to be able to add and remove questions dynamically, database tables were set up to store these custom mappings.

Where a user is applying to a council which makes explicit data-related demands at the funding stage⁷ the user is presented with the DCC clauses which correspond most closely; by answering the DCC clauses, the user *de facto* meets the funder’s requirements. Where a user is applying to a funding council that

⁶ http://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMP_checklist_v2.2_100106-publicVersion.doc

⁷ At end-March 2010, these were: AHRC; BBSRC; ESRC; MRC; Wellcome.

does not make explicit data-related demands at the application stage, the user is presented with a superset of all of the clauses which the mapped funders require, from which the user can add or remove as desired.

At the application (pre-funded) stage, the user interface comprises four columns: the funder's requirements, the equivalent DCC clauses, user input boxes, and a fourth column giving guidance and helpful links. Post-funding, the first of these columns disappears to allow more room on the screen.

An elegant interface using the jQuery Javascript/Ajax library allows the quick addition and removal of questions, and users also have the ability to export their plans as PDF files, which present information in a similar way to the onscreen interface.

4 Testing of DMP Online

The DCC is currently providing dedicated support for the Joint Information Systems Committee (JISC)'s Managing Research Data programme.⁸ Many of the projects within this programme intend to support researchers with Data Management Plan requirements. Several have already consulted the DCC's policy and data management resources⁹ and have volunteered to test DMP Online once the beta version is released in April 2010.

5 Conclusion

We have built a customisable online DMP template tool, into which researchers can enter their own information via an interactive Web interface, depending on their own needs and the requirements of their chosen funder. Users are able to include and exclude individual clauses according to their specific needs, and export their plans in PDF format. Onscreen guidance and suggestions for further help are provided. In time it is hoped that users will be able to view and adapt examples and expressions of good data management practice via an openly accessible library corresponding to each section.

References

- SH. Higgins, S.: The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3(1) (2008), <http://www.ijdc.net/index.php/ijdc/article/view/69>
- SJ. Jones, S.: A Report on the Range of Policies Required for and Related to Digital Curation, v. 1.2. (2009) http://www.dcc.ac.uk/sites/default/files/documents/reports/DCC_Curation_Policies_Report.ipdf
- LL. Lyon, E.J.: Dealing with Data: Roles, Rights, Responsibilities and Relationships (2007), <http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19>

⁸ <http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx>

⁹ <http://www.dcc.ac.uk/resources/policy-and-legal/>

DiLiA – The Digital Library Assistant

Kathrin Eichler, Holmer Hensen, Günter Neumann, Norbert Reithinger,
Sven Schmeier, Kinga Schumacher, and Inessa Seifert

DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
`{firstname.lastname}@dfki.de`
<http://www.dfki.de>,
<http://dilia.b.dfki.de/>

Abstract. In this paper we present the digital library assistant (DiLiA). The system aims at augmenting the search in digital libraries in several dimensions. In the project advanced information visualisation methods are developed for user controlled interactive search. The interaction model has been designed in a way that it is transparent to the user and easy to use. In addition, information extraction (IE) methods have been developed in DiLiA to make the content more easily accessible, this includes the identification and extraction of technical terms (TTs) – single and multi word terms – as well as the extraction of binary relations based on the extracted terms. In DiLiA we follow a hybrid information extraction approach – a combination of metadata and document processing.

1 Introduction

Although the content of digital libraries is growing rapidly, popular portals for digital libraries, such as Google Scholar, Citeulike, ACM digital library still limit the search options to a small set of meta labels (such as author, title, etc.) and only provide a limited text-based search interface. So far, these portals do not use any elaborated visualisation techniques for presenting the search results. This is problematic in two ways. Firstly, since the search options are restricted to metadata, a search query that is not specific enough will easily lead to a long list of search results. Secondly, since no elaborated visualisation techniques are used, navigating through the search result is difficult and time consuming. The goal of DiLiA is to go beyond this level of information access. We especially target users that want to interactively explore the content of the digital library, for example, users that want to investigate a new research area. The DiLiA demonstrator is based on real data in the computer science domain. The database contains 1.2 million abstracts with corresponding metadata from DBLP.

2 Visualisation

The development of the user interface has been led by the design principle that the visual representations should provide clues: what can be done next and what are the possible directions for further search [1]. The user interface consists of a

relational view, visualising relations between the search queries the user specifies; a *hyperlink activated list of search results* with detailed information on each item; and *tools* (e.g., bar charts) for a flexible analysis of the search results.

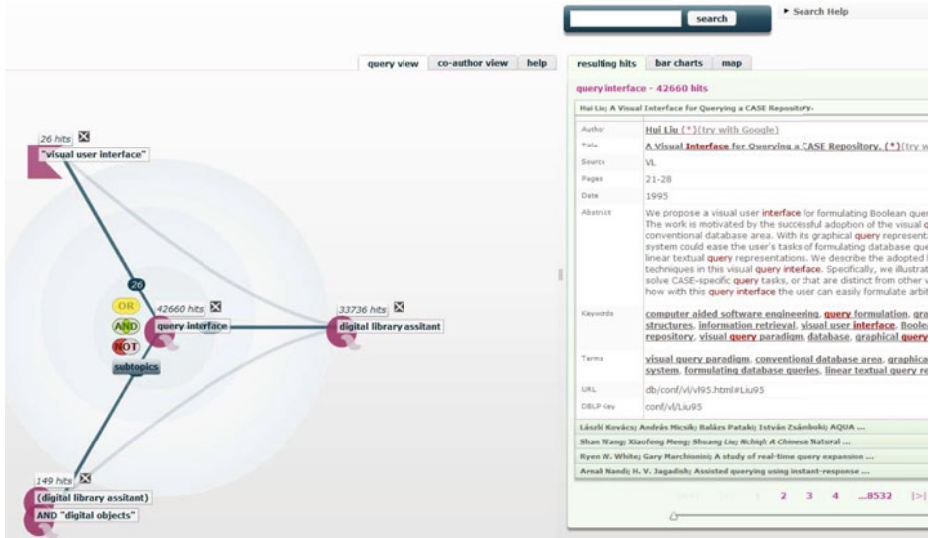


Fig. 1. User interface of the DiLiA demonstrator

Figure 1 shows the user interface (UI) of the DiLiA demonstrator. On the right side of the UI the search panel with the search result list is located. Selecting an item in the result set shows its metadata. Users have different possibilities for stating a search query. The search panel allows the user to enter a search term for searching in the digital libraries metadata. In addition, the user can add items from the hyperlink activated search result list, e.g., a keyword or an author name. On the left side of the UI relations between search queries are displayed (TopicView) in form of a graph (TopicGraph). Each search topic is represented as a TopicBlob (node). Edges between the TopicBlobs show the number documents common in both TopicBlobs. The TopicBlobs can be combined interactively via drag-and-drop on boolean operators (AND, OR, NOT) included in the TopicView (for details see [2]). For each TopicBlob in focus, subtopics can be viewed and selected. The subtopics are automatically generated by dynamically clustering the document abstracts using the Carrot² Clustering engine [1]. On the right side of the UI the user can also switch to various views, supporting visual analytics on the data. The bar chart view shows how many documents have been published in a specific year for the selected topic. Depending on the curve that the bar chart forms the user might be able to see if a research topic is a hot topic or if few papers have been published on the topic lately. The user has also

¹ <http://project.carrot2.org/>

the possibility to use a heatmap view. The heatmap shows using a world map the origin of the publications about the topic and how it emerged over time and enables the user to see where in the world a research topic started and how it spread. On the left side of the UI the user can switch to an author graph, showing for a selected publication the author, the co-authors and the publications of author and co-author and to navigate further.

3 Information Extraction

The goal of IE in DiLiA is on one hand to support digital libraries in the process of making available new material and on the other hand to support users in interactively exploring the content. We have developed a Generalised Name Recogniser (GNR) for identifying domain independent, fully automatically and unsupervised, multi-word technical terms, cf. [3]. Processing only the abstracts of the documents, the current prototype contains these technical terms as automatically generated list of keywords. Based on the identification of TTs in the whole document, we are currently working on unsupervised relation extraction methods. The extracted relations can be used for advanced search and also serve as basis for clustering similar relations/documents. For identifying the TTs [2] we used the nominal group (NG) chunker of the GNR, but the output was modified. For example, coordinated phrases had to be split or text in parenthesis had to be processed separately [3]. Since not every NG is a TT, we needed to find a way to filter the NGs. Inspired by Luhn’s findings [4], who suggested that mid-frequency terms are the ones that best indicate the topic of a document, frequency scores for all NGs using the Live Search API from Microsoft are retrieved. The NGs are then filtered using an upper and lower threshold. We found out that the upper threshold is domain dependent. For computer science documents the best F-measure was achieved with a threshold of 20 mio., for biology 6.5 mio. The extracted TTs serve as the basis for relation extraction.

Fig. 2 shows the information flow. For the IE process all documents are first split into sentences. The identified TTs are then replaced in each sentence with a termID. Three different binary relation strategies have been implemented and are currently being evaluated. The first strategy “surface patterns” is inspired by [5] and uses the following pattern `<TermID1>string<TermID2>` to match each sentence against. For “Verb relations” and “Skeletons” the modified sentences are parsed with the Stanford Parser with dependency tree output. In the “Verb relation” IE method the verb node and direct neighbour nodes containing TTs are extracted. In the “Skeleton” approach [6] the relation consists of information collected by going up the dependency tree starting from pairs of TTs and ending at a common root node.

² An evaluation on a hand-annotated computer science corpus (DBLP) showed that 68.2% of the NGs were identified completely and 31.3% partially (caused by missing prepositional postmodifiers, additional premodifiers and appositive constructions).

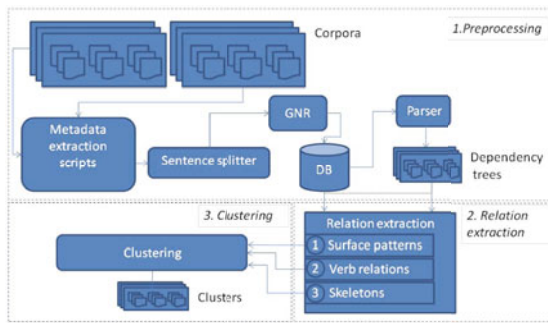


Fig. 2. Information extraction data flow in DiLiA

4 Conclusion and Future Work

In this paper we presented the DiLiA demonstrator³, which provides a novel user interface for interactively navigating in a digital library database. The system also integrates IE methods (automatic extraction of technical terms and binary relations). Currently, we are working on the implementation of the DiLiA system for a Touchmaster touch table (2 x 1.10m) and investigate clustering algorithms for very large data sets.

Acknowledgment

The research project DiLiA is co-funded by the European Regional Development Fund (ERDF) in context of Investitionsbank Berlin's ProFIT program under grant number 10140159. We gratefully acknowledge this support.

References

1. Marchionini, G.: Information-seeking strategies of novices using a full-text electronic encyclopedia. *J. Am. Soc. Inf. Sci.* 40(1), 54–66 (1989)
2. Seifert, I., Kruppa, M.: A pool of topics: Interactive relational topic visualization for information discovery. In: Huang, M.L., Nguyen, Q.V., Zhang, K. (eds.) *Visual Information Communication*. Springer, Heidelberg (2010)
3. Eichler, K., Hensen, H., Neumann, G.: Unsupervised and domain-independent extraction of technical terms from scientific articles in digital libraries. In: Mandl, T., Frommholz, I. (eds.) *Proc. of the Workshop "Information Retrieval"*, Organized as part of LWA, Darmstadt, Germany, September 21-23 (2009)
4. Luhn, H.P.: The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2, 157–165 (1958)
5. Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., Ishizuka, M.: Unsupervised Relation Extraction by Mining Wikipedia Texts Using Information from the Web. In: *Proc. of ACL and the 4th IJCNLP of the AFNLP*, Suntec, Singapore, pp. 1021–1029 (2009)
6. Wang, R., Neumann, G.: Recognizing textual entailment using a subsequence kernel method. In: *Proc. of AAAI-2007*, Vancouver, Canada (2007)

³ Online accessible via: <http://dilia.b.dfki.de/>

Xeproc©: A Model-Based Approach towards Document Process Preservation

Thierry Jacquin, Hervé Déjean, and Jean-Pierre Chanod

6 chemin de Maupertuis 38240 Meylan France
{Firstname.Lastname}@xrce.xerox.com

Developed in the context of the EU Integrated Project SHAMAN, Xeproc© technology lets one define and design document processes while producing an abstract representation that is independent of the implementation. These representations capture the intent behind the workflow and can be preserved for reuse in future unknown infrastructures. Xeproc© is available under Eclipse Public Licence.

1 Xeproc© in the Context of Digital Preservation

Xeproc© was developed in the context of the Integrated Project SHAMAN (<http://shaman-ip.eu/>), co-funded by the European Union within the FP7 Framework. SHAMAN aims at developing a long-term digital preservation framework and tools to analyse, ingest, manage, access and reuse digital objects.

In SHAMAN, Xeproc© use focused on metadata extraction processes [1] operated in the preingest phase. Those processes have been applied to two major types of collections, the Deutsche Nationalbibliothek (DNB) collection of electronic PhD theses (available in PDF format) and digitized collections provided by the Göttinger Digitalisierungszentrum (including proceedings and journals). More precisely, the extracted metadata are produced by XML document processing pipelines dedicated to document structure analysis [2, 3, 4]. Eventually processes developed with Xeproc© have been exported and deployed on an iRODS data grid (<https://www.irods.org>) [5] and the extracted metadata exploited through Chehsire3 (<http://www.cheshire3.org/>) in support of advance search and navigation [6].

The extracted metadata is stored externally to the document themselves, and can be seen as digital objects to be preserved on their own altogether with persistent document ID to enable preservation management and reuse of the metadata. In this view, the metadata extraction processes belong to the context of production of the metadata. By enabling the preservation of logical descriptions of those processes the Xeproc© methodology provides the ground for documenting the metadata provenance information, i.e. information that documents the history of the Content Information [7], where the content is the metadata in this case.

This will support the long term understanding of the metadata and of the extraction processes and will enable their reconstruction as technology evolves and improves over time.

More specifically, within the context of SHAMAN and digital preservation, Xeproc© models XML pipelines and XML validation checkpoints. These capture the intent behind the workflow irrespective of the implementation at a given point in time. These abstract representations are preserved, so that the Xeproc© models can be seen as independent specifications to be instantiated and deployed over time and as technology evolves. These logical and persistent descriptions, when associated with the accurate components, are interpreted or translated into any SOA orchestration language to produce logically structured documents (typically XML). These make explicit how the source document content is logically and semantically organized.

2 The Xeproc© Technology

Xeproc© technology can be used to build a wide range of applications based on document processing, including transformation, extraction, indexing and navigation. It can be easily integrated with more global business processes and customized to match specific requirements and infrastructures. In the spirit of service-oriented architecture (SOA), Xeproc© embeds references to services and documents and provides loose coupling not only to services but also to data resources, with respect to both their location and format.

Available on Eclipse 3.5.1 under the Eclipse Public License, Xeproc© combines a domain-specific language (DSL), an associated graphic designer and extension APIs (application programming interfaces).

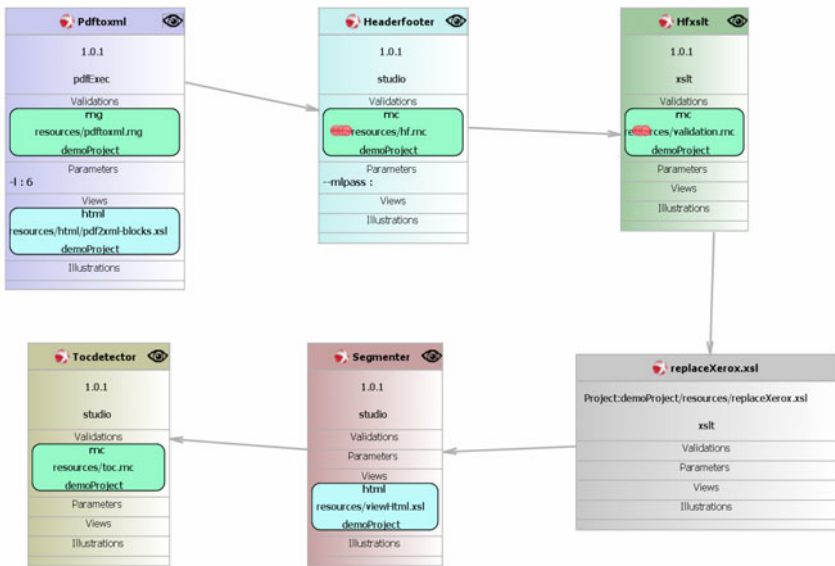


Fig. 1. A Xeproc© template under design

3 The Xeproc© DSL: Extensible, Easy to Use and Focussed

The Xeproc© Domain-Specific Language (DSL) is used to describe the document process one wants to design. It specifies a chain of processing steps, which may point to components such as document services or project-specific resources. All components take a document as input and generate another document as output.

To take full advantage of Xeproc©, the designer links processing steps with validation resources. While validations are traditionally exploited just before deployment, the Xeproc© Designer is conceived in such a way that they are exploited throughout the design process. Thanks to a continuous monitoring mechanism, validations not only verify but also specify, and lead the design process from the specification to instantiation.

In addition, processing steps can be linked to visualization specifications, highlighting selected outputs. These views, which are captured on demand and throughout the entire monitoring of the process, make it easier to identify and pinpoint errors, undertake corrections or consult the relevant experts.

The Xeproc© DSL is open enough to support any document format, validation syntax and resource location. The Xeproc© DSL is defined by a dedicated XML schema available at <http://www.xrce.xerox.com/Xeproc>.

4 The Xeproc© Graphic Designer

The Xeproc© Graphic Designer is a user-friendly Eclipse plug-in editor which allows the user to manipulate abstract representations of objects relevant to the Xeproc© application domain.

The Designer provides an intuitive representation of underlying Xeproc© models and the ability to draw, rearrange and tune document-processing chains. This is achieved by combining project-specific resources (processing components, validations and views) with generic document services organized in a palette. The processing elements are represented as boxes, intermediate documents as arrows and validation constraints and views as icons on boxes.

The Designer was generated from the Xeproc© model using the EMF/GMF (Eclipse Modelling Framework and Graphical Modelling Framework) technologies provided by Eclipse (<http://www.eclipse.org/>). Model-Driven Architecture methodologies [8] supported by the Object Management Group (<http://www.omg.org/>) were applied.

4.1 Example Scenario

A document transformation project will typically create an Eclipse project, share it amongst all the technical partners and initialize it with the reference resources such as documents, requirements and schemas to be validated. The process designer will consider the context and customize the palette of components with those considered useful from a site update. From there (s)he will start the building process and may drag and drop from the component palette or from the project workspace, quickly drawing specific logical and persistent pipelines for document analysis and transformation.

Links: <http://www.xrce.xerox.com/Xeproc>

Acknowledgements

This work is supported by the Large Scale Integrating Project SHAMAN, co-funded under the EU 7th Framework Programme (<http://shaman-ip.eu/>).

References

1. Dobрева, M., Kim, Y., Ross, S.: Designing an automated prototype tool for preservation quality metadata extraction for ingest into digital repository. *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies 5* (2008)
2. Déjean, H., Meunier, J.-L.: Logical document conversion: combining functional and formal knowledge. In: *ACM Symposium on Document Engineering 2007*, pp. 135–143 (2007)
3. Déjean, H., Meunier, J.-L.: On tables of contents and how to recognize them. *International Journal on Document Analysis and Recognition, IJDAR* 12(1), 1–20 (2009)
4. Mao, S., Rosenfeld, A., Kanungo, T.: Document structure analysis algorithms: a literature survey. In: Kanungo, T., Smith, E.H.B., Hu, J., Kantor, P.B. (eds.) *Document Recognition and Retrieval X*, vol. 5010, pp. 197–207. SPIE, San Jose (2003)
5. iRODS: integrated Rule Oriented Data System. White Paper. Data Intensive Cyber Environments Group. University of North Carolina at Chapel Hill, University of California at San Diego
6. Sanderson, R., Larson, R.: Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval. In: *JCDL* (2005)
7. Reference model for an Open Archival Information System (OAIS). Blue Book CCSDS 650.0-B-1, Consultative Committee for Space Data Systems (2002), <http://public.ccsds.org/publications/archive/650x0b1.pdf>
8. <http://www.omg.org/mda/>

A Prototype Personalization System for the European Library Portal

Marielena Kyriakidi, Lefteris Stamatogiannakis, Mei Li Triantafyllidi,
Maria Vayanou, and Yannis Ioannidis

Department of Informatics and Telecommunications, MaDgIK Lab
University of Athens, Panepistimioupolis, Ilissia, 15784 Athens, Greece
{marilou, estama, meili, vayanou, yannis}@di.uoa.gr

Abstract. In this demonstration, we present a flexible system that enables the provision of personalized functionalities to digital libraries. The system has been developed based on the needs of *The European Library* portal and will be demonstrated in that particular context, but could be applied more generally. It implements a broad set of data processing, analysis, and mining techniques over the portal's log files, using an environment called *madIS*. It enables on-line result contextualization and adaptation through the development of REST web services, which are responsible for retrieving and appropriately integrating the extracted information. The demonstration also features a web-based visualization tool for showing the output of the log analysis performed.

Keywords: Log mining, pattern extraction, profiling, personalization.

1 Introduction

The European Library (TEL) portal offers access to the resources of the 48 national libraries of Europe. In the TEL home page, users can initiate simple keyword searches within subsets of library resources, referred to as *collections*. Users may search in a pre-selected set of collections or select particular collections in a customized fashion. In the results page of a query, the relevant collection list is placed on the left and the documents from each individual collection are placed on the main panel.

To personalize this functionality, i.e., customize it to the profile of individual users or groups, we have studied the portal characteristics and have analyzed the TEL usage logs. For example, since query results are grouped per collection, instead of being fused into a unified ranked list, correct use of collection selection features is crucial for users to effectively exploit TEL services and content. Log analysis results, however, show that in almost 65% of sessions, users perform no collection specification but search within the default collection. Likewise, although login functionality is provided for user registration, these are hardly used, imposing a significant obstacle to the extraction of accurate personal profiles.

The above and other important findings of our investigation have formed the basis for the services offered by the Personalization Prototype to be demonstrated. In particular, to address the data sparsity observed, collaborative approaches are employed,

through the specification of group-level user models. Also, usage analysis revealed that users from the same country tend to exhibit several commonalities regarding their preferences [3], leading the way for the definition of a National user group. In addition to the Personal and National levels, a Global profile exploits the “wisdom of the crowds”.

Moving in these directions, we have developed a prototype Personalization system that provides personalized and collaborative functionality to TEL portal. It is flexible and easily adaptable to portal changes and new applications requirements, while it can be quickly integrated into other digital library portals as well.

2 System Architecture

Figure 1 depicts the main components of the personalization system, as integrated with TEL. User requests are issued to the portal, which is hosted in TEL server. The results are returned to the user, while all user interaction is recorded in log files.

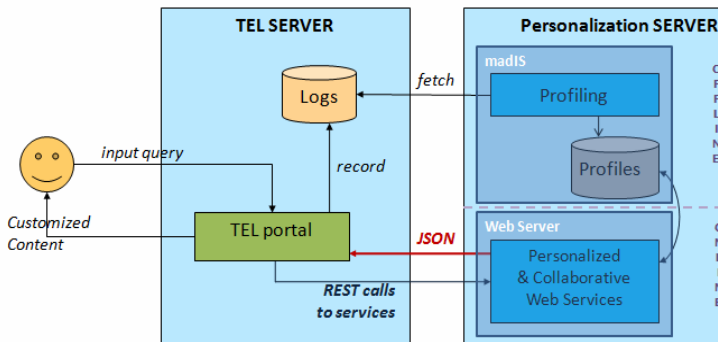


Fig. 1. System Architecture

To adapt the results generated to personal and context-based information, we have followed an implicit feedback approach, using various techniques to process and mine the log data to extract several useful patterns and user preference profiles. All knowledge extraction has been captured by a set of workflows that are executed by the *madIS* prototype processing and analysis environment [2] and are stored in the Personalization server under a relational schema. Due to the high computational complexity of profiling and to avoid any degradation of the system’s performance, profile extraction takes place offline, in compliance with the basic principle of online / offline separation [1]. The profiling service is executed periodically for fetching new log data, which is subsequently used for updating existing patterns and profiles.

The information extracted is accessed at run-time by a set of *REST* web services that are responsible for combining related patterns and profiles to provide the corresponding customized information. They use *madIS* for data processing and retrieval and their results are returned to the invoking entity, encoded in *JSON*. These services are also hosted in the Personalization server and web service functionality and protocols are provided by the *Tornado* web server.

3 Log Mining and Profiling

TEL usage mining has been implemented through madIS workflows, containing a series of queries expressed in an SQL-based declarative language extended with User Defined Functions (UDFs) implemented in Python. As in traditional data mining, workflows include activities for Data Collection, Log Cleaning, Log Transformation, and Pattern Extraction.

In Data Collection, the “Initialization workflow” imports and integrates various types of TEL data (e.g., application log files, collection descriptions, users’ registration information, saved queries, and favorites) as well as external data (e.g. GeoIP database, ISO country codes, stop word lists, and statistical language models).

In Log Cleaning and Transformation, several workflows resolve inconsistencies and assemble data into an integrated and comprehensive view. For example, a typical web usage mining activity is *search session reconstruction*, grouping user actions in comprehensive efforts towards one goal. One workflow employs the popular 30 minutes of inactivity timeout, which is shown to be both effective and efficient [4]. Another workflow maps sessions to country codes, which is useful for subsequently extracting national patterns. Finally, another workflow classifies sessions into Expert or Non-Expert (based on ad-hoc session characteristics found critical during log analysis), which is useful for subsequently emphasizing expert behavior more heavily.

In Pattern Extraction, five additional workflows are used to construct specialized patterns or profiles. Depending on the goal of each workflow, it may also include additional processing, e.g., stop word removal, stemming, and language detection. An Apriori-like data mining algorithm has been implemented for extracting correlations among query keywords and is applied at three profile levels: Personal, National, and Global. Expert sessions are emphasized within the computation of group profiles using heuristically defined weights. In addition, two term-indexing table structures are constructed based on query-term frequency for “original query recommendation”.

Regarding collection usage, *freqency* metrics are employed, combining frequency and recency, to effectively capture concept drift and temporal trends. Computation is performed again at all profile levels, resulting in three ranked lists of collections. Moreover, correlations between queries and collections are extracted over the group-level profiles, based on frequency measures, while some additional statistics are computed to quantify secondary user actions, such as selection of Advanced Search Fields, Collection Themes, etc. Finally, a user similarity matrix is constructed capturing similarity between each pair of users over a variety of dimensions (user interests, collection usage, queries, favorite object descriptions) that are integrated into a unified similarity score.

4 Demonstration Overview

Addressing the needs and characteristics of the TEL portal, the following five adaptive services are provided, which combine all available profile levels while always emphasizing the finer grained ones:

1. *Personalized Collection Ranking* computes a personalized and context-aware ranking of TEL Collections

2. *Collaborative Query Suggestion* produces “original query recommendations”
3. *Personalized Term Suggestion* generates a related terms cloud
4. *Collaborative Query based Collection Recommendation* generates a list of top recommended collections with regard to the query issued
5. *Personalized User Notification* retrieves top similar users along with their preferred content, thus providing an enhanced, user-aware platform.

A demonstrator tool has been developed, receiving the *JSON* response of each service and presenting it within a web browser window. The graphical interface depicted in Figure 2 has been employed for testing and experimentation during log analysis and it has been extended for demonstrating key statistical results and patterns. During the demonstration, users will be able to set a variety of input parameters and explore the output results of each service using the desired graphical representations. The interface is targeted towards results exploration and it is not meant to be used by TEL end users.

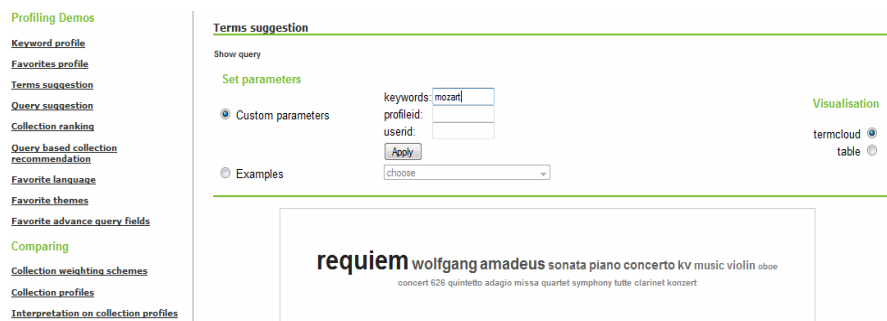


Fig. 2. Graphical Interface of Demonstration Tool

Acknowledgments. This work was done within the TELplus project (www.theeuropeanlibrary.org/telplus/), funded by the European Commission. We would like to thank the TEL Office for its valuable help during the project, especially Georgia Angelaki and Anna Gos. We are also grateful to our colleagues from the Univ. of Padua, especially Maristella Agosti and Giorgio Maria Di Nunzio for their feedback and valuable discussions during our collaboration within the project.

References

1. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. *Communications of the ACM* 43(8), 142–151 (2000)
2. <http://code.google.com/p/madis/>
3. Agosti, M., Crivellari, F., Di Nunzio, G.M., Ioannidis, Y., Stamatogiannakis, E., Triantafyllidi, M.L., Vayanou, M.: Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for The European Library. In: *IRCDL*, pp. 120–135 (2009)
4. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (2005)

Meta-Composer: Synthesizing Online FRBR Works from Library Resources

Michalis Sfakakis, Panagiotis Staikos, and Sarantos Kapidakis

Laboratory on Digital Libraries & Electronic Publishing,
Department of Archive & Library Sciences, Ionian University
Plateia Eleftherias, Paleo Anaktoro GR-49100 Corfu, Greece
{pstaikos, sarantos}@ionio.gr

Abstract. Next generation display and indexing of cataloging records are mainly influenced from the development of the FRBR conceptual model. While the process for collecting all relevant bibliographic records in a catalogue to an FRBR work entity has been extensively developed and tested in non interactive (offline) environment, the corresponding process has not been explored when meta-searching. This work presents the implementation and use of alternative clustering algorithms and similarity metrics for the composition of the FRBR work entities in the configurable meta-search engine *meta-Composer*. Moreover, it introduces a tool for the evaluation of the composition methods, which can be used either as complementary to the configuration process for the use of the best clustering methods to the searched catalogues or as a general testbed for the evaluation of the FRBR work entities composition process.

1 Introduction

After the development of the Functional Requirements for Bibliographic Records (FRBR) from IFLA [5], current online search interfaces have been criticized for their inability to find and collocate all versions of a distinct intellectual work represented by many bibliographic records [15], while Mimno et al. have explored the benefits of moving to hierarchical catalogues for searching and browsing [9].

Running projects for FRBRizing the display and indexing of cataloguing are focusing on either the implementation of the FRBR conceptual model or the development of tools and methods for synthesizing existing records to FRBR entities. For the composition of the FRBR entities, all tools generate key identifiers for every entity, while the complexity of the algorithm depends on the final goals of the tool. Depending on the entity, the key may consist of more than one part (subkeys), while many different fields or subfields may be used for the key construction. It is worth mentioning that the existing methods emphasize on the selection and the preprocessing of the data fields, while they mainly apply exact key comparison procedures and rarely use string similarity techniques.

A few tools for FRBR entities composition already exist such as the FRBR display tool made available by the Library of Congress Network Development and MARC Standard office [8] as an open source XSLT program. The tool is based on the analysis made on how to use FRBR model for clustering retrieved records in a more

meaningful display [7]. Other tools, mainly intended to catalogue conversions, are the work-set algorithm developed by the OCLC [10] and the tool developed as part of the Norwegian BIBSYS FRBR project for converting the BIBSYS bibliographic database [1]. Freire et. al. [4] present an experiment using string similarity techniques for the identification of FRBR work entities in a library catalogue. More specifically, they have shown that similarity metrics can be used in the process of FRBR works identification within bibliographic catalogues resulting to low rate of error for both mismatches and missed matches.

meta-Composer [12] is a meta-search engine that composes work level entities for display and avoids query failures by substituting the unsupported Access Points in the context of the Z39.50/SRU [3] environment. *meta-Composer* is developed at the Laboratory on Digital Libraries & Electronic Publishing at the Ionian University and the running version (available at <http://dlib.ionio.gr/metacomposer>) meta-searches the following sources: *Library of Congress* (US), *Library and Archives Canada*, *MELVYL*, *COPAC Academic & National Library Catalogue* (UK), *Hellenic Academic Libraries Union Catalogue* and *University of Crete* (Greece).

This work reports to the following areas: (i) the adaptation and the implementation of various clustering techniques in a toolkit, combined with the use of a number of inter-cluster and entity key similarity metrics for the composition of FRBR entities, (ii) the development of the *frbrCluster* service for the application and the evaluation of the implemented clustering techniques on any given bibliographic record set conforming to any of the MARC21, UNIMARC and MODS standard and finally, (iii) the extension of *meta-Composer* in the work entities composition procedure, by selecting and using any of the implemented clustering techniques and similarity metrics from the toolkit.

2 Applying Clustering Techniques

A number of clustering algorithms have been proposed and used from many disciplines for several decades [14]. According to the commonly referred to as hierarchical and partitional categories of the clustering algorithms [6], the bottom-up *Hierarchical Agglomerative Clustering* (HAC) and the top-down *Bisecting K-means* algorithms were implemented from the hierarchical category, while the *Single Pass* algorithm was implemented from the partitional category.

As we already mentioned, the entity key identifiers used for the matching process during the cluster composition may contain sub-keys. Thereafter, the algorithms were implemented to handle such complex keys by running recursively the process for all subkeys and also to apply them different similarity metrics. For measuring the string similarity, representative metrics, token or character based [4], were implemented.

The development of the clustering toolkit was driven by the need to develop a general platform upon which various clustering methods and string similarity metrics can be tested and evaluated. In our context, the goal is to provide a testbed for the implemented clustering techniques for the FRBR work entities composition on any given bibliographic record set conforming to any of the MARC21, UNIMARC and MODS standard. For the overall evaluation of the implemented clustering algorithms in

the toolkit, the measures *F-Measure*, *Cluster Purity* and *Entropy* [2, 16] are also provided. The core functionality of the FRBR clustering toolkit is available for use to any other system through a C language interface, while the web based *frbrCluster* service exposes toolkit's functionality to any ordinary user.

Applying clustering in a meta-search context is close to *query clustering*, hence the highly variable nature of the returned result sets and the tight efficiency constraints are under consideration [2]. Therefore, for the overall performance of the meta-search engine a good balance between the efficiency and the effectiveness of the clustering techniques is required. Moreover, *meta-composer* receives and processes the resulting records gradually, without having to wait for all sources to respond, while a fast response to the user is a key issue for the acceptance of the system. Thereafter, in order to achieve good performance and the system to be able to present partial results to the user all the implemented algorithms are adapted to process the incoming records also incrementally.

meta-composer utilizes the clustering facilities via the C language interface. Among the alternative implemented clustering algorithms and similarity metrics, the running version is configured to apply the incremental form of the *HAC* algorithm with complete linkage method for inter-cluster similarity. Given that the FRBR work entity key is composed from the *author* and the *title* subkeys, the jaro-winkler character based similarity metric with threshold 0.90 is selected for the *author* subkey and the cosine similarity with TF-IDF values with threshold 0.70 is selected for the *title* subkey.

3 Discussion

A toolkit consisting of a number of different clustering techniques and string similarity metrics for the FRBR work entities composition has been implemented and is publicly available at the Laboratory on Digital Libraries & Electronic Publishing at the Ionian University. The functionality of the toolkit is demonstrated from two alternative services. *frbrCluster* (<http://dlib.ionio.gr/frbrcluster>) demonstrates the use of the clustering techniques on any given bibliographic record set conforming to any of the MARC21, UNIMARC and MODS standard and can be used as a testbed and evaluation system. Alternatively, *meta-Composer* (<http://dlib.ionio.gr/metacomposer>) is a configurable running meta-search service using the implemented clustering tools for displaying work-centric results to the user. Furthermore, the incremental version of the clustering algorithms enables *meta-Composer* to inform the user for the current state of the clustering process with the incoming records at any time of the retrieval task.

Some preliminary results from the *frbrCluster* application on small data sets containing bibliographic records, where the FRBR *work* clusters were manually constructed, show that the *Single Pass* algorithm was the fastest one and performs very close to the *Bisecting K-means*. When the algorithm processes the records incrementally, the *HAC* reduces its execution time dramatically without significant degradation in the quality of the clusters. In general, the character-based similarity metrics need more execution time than the token-based. The token-based similarity metrics perform better when low threshold is set. In contrast, the character-based metrics perform better when high threshold is used.

The use of a more reliable data set with more bibliographic records and different work cases is in our future plans. Moreover the enrichment of the tool with more clustering algorithms and similarity metrics will further improve the usability of the developed tool.

References

1. Aalberg, T., Haugen, F., Husby, O.: A Tool for Converting from MARC to FRBR. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 453–456. Springer, Heidelberg (2006)
2. Andrews, N., Fox, E.: Recent Developments in Document Clustering. In: Department of Computer Science, Virginia Tech, Blacksburg, VA (2007), <http://eprints.cs.vt.edu/archive/00001000/01/docclust.pdf>
3. ANSI/NISO: Z39.50 Information Retrieval: application service definition and protocol specification: approved May 10 (1995)
4. Freire, N., Borbinha, J., Calado, P.: Identification of FRBR Works Within Bibliographic Databases: An Experiment with UNIMARC and Duplicate Detection Techniques. In: Goh, D.H.-L., Cao, T.H., Sølvyberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 267–276. Springer, Heidelberg (2007)
5. IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records: Final Report, UBCIM Publications-New Series, vol. 19. K. G. Saur, Munchen (1998), <http://www.ifla.org/VII/s13/frbr/frbr.htm>
6. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, Chichester (1999)
7. Library of Congress Network Development and MARC Standards Office: Displays for Multiple Versions from MARC 21 and FRBR (2001), <http://www.loc.gov/marc/marc-functional-analysis/multiple-versions.html>
8. Library of Congress Network Development and MARC Standards Office: FRBR Display Tool Version 2.0 (2004), <http://www.loc.gov/marc/marc-functional-analysis/tool.html>
9. Mimno, D., Grane, G., Jones, A.: Hierarchical Catalog Records: Implementing a FRBR Catalog. D-Lib Magazine 11(10) (2005), <http://www.dlib.org/dlib/october05/crane/10crane.html>
10. OCLC: FRBR work-set algorithm (2005), <http://www.oclc.org/research/software/frbr/default.htm>
11. Sfakakis, M., Kapidakis, S.: Eliminating Query Failures in a Work Centric Library Meta-Search Environment. Library Hi Tech. 27(2), 286–307 (2009)
12. Tillett, B.: What is FRBR? A Conceptual Model for the Bibliographic Universe, <http://www.loc.gov/cds/downloads/FRBR.PDF>
13. Xu, R., Wunsch II, D.C.: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks 16(3), 645–678 (2005)
14. Yee, M.: FRBRization: a Method for Turning Online Public Finding Lists into Online Public Catalogs. Information Technology and Libraries 24(3), 77–95 (2005)
15. Yoo, I., Hu, X.: A Comprehensive Comparison Study of Document Clustering for a Bio-medical Digital Library MEDLINE. In: JCDL 2006, Chapel Hill, North Carolina, USA, June 11–15 (2006)

Digital Library in a 3D Virtual World: The Digital Bleek and Lloyd Collection in Second Life

Rizmari Versfeld¹, Spencer Lee², Edward Fox², Hussein Suleman¹, and Kyle Williams¹

¹ Department of Computer Science
University of Cape Town
55A Polo Road
Observatory, Cape Town
South Africa
(+27) 0733416692

{vrsriz001, hussein, kwilliams}@uct.ac.za

² Department of Computer Science
Virginia Tech, M/C 0106
Blacksburg, VA 24061 USA
+1-540-231-5113
{zamfir, fox}@vt.edu

Abstract. This research explores and demonstrates the process of setting up a 3D representation of a typical web-based digital library called ‘The Digital Bleek and Lloyd collection (lloydbleekcollection.cs.uct.ac.za)’ in the popular 3D virtual world, ‘Second Life’. The processes of building, scripting, and evaluation of the 3D exhibit are discussed. The report concludes that SL is a good platform for this kind of cultural representation. At a university level it could be used to showcase and share researchers’ work.

Keywords: Second Life, virtual worlds, 3D, Digital Libraries, Bleek and Lloyd, Bushman heritage.

1 Introduction

The Digital Bleek and Lloyd is a collection of scanned notebooks and illustrations documenting the Southern African Bushman culture. More specifically the notebooks contain words and stories written by Wilhelm Bleek, Lucy Lloyd, and Jemima Bleek in the !xam and !kun languages. All the drawings and watercolours of the Bushmen !han#kass’o, Dia!kwain, Tamme, luma, !nanni and Da are included in the collection [1].

Second Life (SL) is a popular general-purpose 3D virtual world where users can create their own objects and program them to perform various tasks using the Linden Scripting Language (LSL). These tasks include displaying and navigating slideshows, videos, and other media, handing out informational note cards, performing AI activities for bots, and much more. Many organizations around the world like NASA, IBM [2], Sun Microsystems [3], the Digital Libraries Federation [4], and JCDL/ECDL 2009 (Joint/European Conference on Digital Libraries) [5] have hosted exhibitions,

seminars, and conference sessions in SL. Many universities have a presence in SL, and some of them even offer courses that students can attend in SL [6] [7].

The goal of this research is to explore whether an existing digital library collection can be re-represented in today’s 3D virtual worlds and to see if the newly created 3D representation can both convey information and effectively engage with people. This research work has been done in collaboration with a Virginia Tech team which is conducting an U.S. NSF (National Science Foundation) funded research project (IIS-0910183) on the topic of digital preservation education in SL.

2 Processes

2.1 Building

A 1600 sq.m. parcel at <http://slurl.com/secondlife/Digital%20Preserve/190/37/22/> was borrowed from the SL region called ‘Digital Preserve’ which is managed by the above-mentioned Virginia Tech team. The avatar ‘Riz Juneberry’ was designated as the owner of the parcel, to create objects (called ‘prims’ in SL) and scripts on the parcel; most development was done using the avatar.



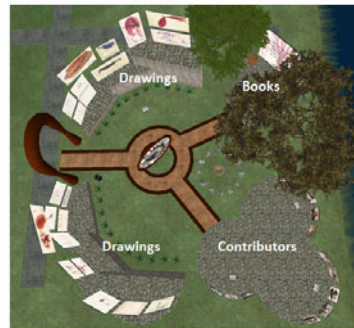
Entrance of the original web-based collection



Entrance of the exhibit in Second Life



Navigation menu of the original collection



Geographical navigation of the exhibit in SL

Fig. 1. Comparison between the original web-based digital Bleek and Lloyd collection and the exhibit in Second Life

The basic building tool provided by SL was used to create most objects found in the exhibit. Primitive shapes were modified and linked together to form display areas, media screens, signs, and other objects. Textures used for the objects are custom-made, from the original collection, or from the inventory library provided by SL. The GIMP Texturize plug-in was used to create textures from normal images. The Bushman illustrations also needed to be cropped and resized in some cases so as to reduce loading times in SL. This was done with the GIMP Batch process plug-in. All image processing was done with GIMP 2.6.4.

Only one animated texture was used in the exhibit. This was the fire texture. The animation was handled by a script that looped through frames in a single image. The frames of the fire animation were devised using particleIllusion 3.0 software that creates 2D particle effects.

The slideshow in the exhibit was created using Microsoft PowerPoint 2007. It was then exported as a series of JPEG images that were uploaded to SL. Navigation of the slideshow is handled by a script. Once all images had been processed and uploaded they could be applied to SL objects. Then the objects were arranged so as to allow for easy navigation.

Figure 1 shows the final outcome of the building process and also the comparison between the original web-based digital collection and the newly created SL exhibit.

2.2 Scripting

The final part of preparing the Bleek & Lloyd exhibit involved scripting objects. Scripts were used for 6 purposes: animation, slideshow navigation, floating text, note card distribution, embedded web links, and note card storage. Floating text was used in various objects to give the user instructions, and also used to identify the Bleek & Lloyd contributors. Certain objects contain embedded web links which open in a browser when the object is clicked on. The display stands in the exhibit used a note card distribution script to give a user a note card when the stand is clicked on. These note cards provide more information about the contributors.

2.3 Evaluation

Once development of the exhibit had been completed, subjects to evaluate the exhibit were recruited. This brief evaluation consisted of navigating the Digital Bleek & Lloyd website first, exploring the SL exhibit, and completing a short online survey. Due to time constraints, only five subjects were recruited. The main 3 questions of the survey were:

1. Do you think the SL exhibit contributes to the online Bleek & Lloyd in a meaningful way?
2. Do you find the exhibit more interesting/engaging than the website?
3. Do you think development of the exhibit or similar project in Second Life should continue?

Figure 2 shows the very encouraging results of the brief evaluation.

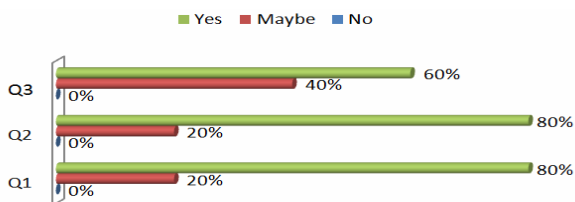


Fig. 2. Answers to multiple choice survey questions

3 Conclusion

The process of creating content in SL is a straightforward, although time consuming work. The scripting functionality of SL objects makes them very flexible and allows for the possibility of very complex and creative scripts to add to the functionality of an SL environment. This makes SL an interesting and, according to the survey results, effective platform for representation of information. SL could be a good educational tool in schools, and tertiary institutions could find SL to be a good way of showcasing current research. It could also be a platform for communication amongst communities of researchers and students interested in the same topics.

References

1. The Digital Bleek and Lloyd, <http://lloydbleekcollection.cs.uct.ac.za>
2. IBM Business Center, <http://www.ibm.com/3dworlds/businesscenter/us/en/>
3. Sun Microsystems Presence In Second Life, <http://www.sun.com/aboutsun/media/presskits/secondlife/>
4. Digital Libraries Federation Fall Forum 2009, <http://www.diglib.org/forums/fall2009/secondlife.htm>
5. Lee, S.J., Fox, E.A., Marchionini, G., Velasco, J., Antunes, G., Borbinha, J.: Virtual DL poster sessions in Second Life. In: Proceedings of the 9th Joint Conference on Digital Libraries, JCDL (2009)
6. Education in Second Life, http://www.simteach.com/wiki/index.php?title=Second_Life_Education_Wiki
7. Ritzema, T., Harris, B.: The use of Second Life for distance education. *J. Comput. Small Coll.* 23(6), 110–116 (2008)

Appendix

Doctoral Consortium, Workshops, Tutorials, and Panel

Retrieving Data from Mind Maps to Enhance Search Applications

Jöran Beel

Otto-von-Guericke University, Computer Science/ITI/VLBA-Lab, Magdeburg, Germany
joeran.beel@ovgu.de

Abstract. Web search, academic search and expert search engines often have difficulties in classifying and ranking objects and generating summaries for them. In this paper I propose that data retrieved from mind maps may enhance these search applications. For instance, similar to anchor text analysis, documents linked in a mind map might be classified with text from the linking nodes. This idea, along with additional ideas for my PhD, related work, the indented methodology, first research results and issues that I would like to discuss with the ECDL doctoral consortium are presented in this paper.

Keywords: Mind maps, information retrieval, search applications, digital libraries, document classification, document relatedness, expert finding.

Using Digital Libraries to Support a Feasibility Evaluation of a Brazilian Metadata to Learning Objects to Web, Mobile and Digital Television Platforms

Júlia Marques Carvalho da Silva and Rosa Maria Vicari

Federal University of Rio Grande do Sul, Porto Alegre, Brazil
julia.silva@bento.ifrs.edu.br, rosa@inf.ufrgs.br

Abstract. This paper presents a study of the feasibility of a Brazilian metadata specification (called OBAA) to learning objects. The specification is focused to attend multiplatform: web, mobile and digital television. To verify the feasibility we built three digital libraries with learning objects using three metadata standards: IEEE LOM, Dublin Core and OBAA. In addition, we want to verify it through a questionnaire applied with teachers.

Keywords: Learning Objects, Metadata, OBAA.

Using Digital Libraries to Support a Feasibility Evaluation of a Brazilian Metadata to Learning Objects to Web, Mobile and Digital Television Platforms

Júlia Marques Carvalho da Silva and Rosa Maria Vicari

Federal University of Rio Grande do Sul, Porto Alegre, Brazil
julia.silva@bento.ifrs.edu.br, rosa@inf.ufrgs.br

Abstract. This paper presents a study of the feasibility of a Brazilian metadata specification (called OBAA) to learning objects. The specification is focused to attend multiplatform: web, mobile and digital television. To verify the feasibility we built three digital libraries with learning objects using three metadata standards: IEEE LOM, Dublin Core and OBAA. In addition, we want to verify it through a questionnaire applied with teachers.

Keywords: Learning Objects, Metadata, OBAA.

Supporting User Selection of Digital Libraries

Helen Dodd

Department of Computer Science, Swansea University, United Kingdom
cshelen@swansea.ac.uk

Abstract. This research aims to support users in identifying collections (e.g. digital libraries) that are authorities on the topic they are searching for. These collections should contain a large proportion and quantity of relevant documents, such that they may serve both current and (related) future information needs. This paper presents our research goals for this search task, and the steps taken thus far to achieve them. In addition, we provide our plans for future research in this area.

Keywords: Digital Libraries, User Selection, Search, Collections.

Measuring Document Relatedness by Citation Proximity Analysis and Citation Order Analysis

Bela Gipp

UC Berkeley, Ovgu, Berkeley, CA, USA
gipp@berkeley.edu

Abstract. This work-in-progress paper gives an overview of my PhD project. It focuses on two new approaches: Citation Proximity Analysis (CPA) and Citation Order Analysis (COA). They were developed to identify related plagiarized/translated documents, respectively, for the purpose of research paper recommender systems, but can also be applied in other fields like analyzing patent specifications, mind maps and in a modified version, for websites. It is also shown that CPA and COA cannot replace text analysis and existing citation analysis approaches for research paper recommender systems since they all have their own strengths and weaknesses, but could deliver the best results when combined.

Keywords: Document Similarity, Relatedness, Clustering, Plagiarism Detection, Duplicate Detection, Citation Analysis, Citation Proximity Analysis, Citation Order Analysis, Language Independent.

Measuring Document Relatedness by Citation Proximity Analysis and Citation Order Analysis

Bela Gipp

UC Berkeley, OvGU, Berkeley, CA, USA
gipp@berkeley.edu

Abstract. This work-in-progress paper gives an overview of my PhD project. It focuses on two new approaches: Citation Proximity Analysis (CPA) and Citation Order Analysis (COA). They were developed to identify related plagiarized/translated documents, respectively, for the purpose of research paper recommender systems, but can also be applied in other fields like analyzing patent specifications, mind maps and in a modified version, for websites. It is also shown that CPA and COA cannot replace text analysis and existing citation analysis approaches for research paper recommender systems since they all have their own strengths and weaknesses, but could deliver the best results when combined.

Keywords: Document Similarity, Relatedness, Clustering, Plagiarism Detection, Duplicate Detection, Citation Analysis, Citation Proximity Analysis, Citation Order Analysis, Language Independent.

User Expectations and Evaluation of Multilingual Information Access in Digital Libraries

Maria Gäde

Berlin School of Library and Information Science, Berlin, Germany
maria.gaede@ibi.hu-berlin.de

Abstract. While the importance of multilingual access to information systems is undoubted, few truly operational systems exist and can serve as examples. This dissertation addresses the issue of what the user expectations and the consequences for system development are in a multilingual information environment. It starts with a general overview over the aspects of multilingual access in digital libraries. Building on previous experiences, the study focuses on a combination of log file analysis and interviews on user needs and desired features for multilingual access based on a functional digital library with multilingual requirements (Europeana). I present the Europeana Clickstream Logger, which logs and gathers extended information on user behavior, and show first examples of the data collection possibilities. The outcome of the analysis is a description of user requirements. The dissertation concludes with the development of a possible approach for the design of multilingual information systems.

Keywords: CLIR, MLIA, user study, Log file analysis.

Semantic Interoperability in Europeana

Marlies Olensky

Humboldt University, Berlin School of Library and Information Science, Berlin, Germany
marlies.olensky@ibi.hu-berlin.de

Abstract. Interoperability of digital libraries has been a research challenge for some time, especially its semantic aspect. In a digital library environment like Europeana, where a vast amount of information objects from heterogeneous sources will be made available, it is important to find standards that enable semantic interoperability. This dissertation aims at evaluating the applicability of CIDOC CRM for providing semantic interoperability in a digital library environment. To achieve this, the current use of CIDOC CRM in digital cultural heritage projects will be examined and its role for semantic interoperability will be determined. Limitations, benefits and prerequisites will influence the answer to the question on how CIDOC CRM can be integrated in Europeana in order to enhance semantic interoperability.

Keywords: Semantic interoperability, digital library, evaluation, CIDOC CRM, Europeana.

Semantic Interoperability in Europeana

Marlies Olensky

Humboldt University, Berlin School of Library and Information Science, Berlin, Germany
marlies.olensky@ibi.hu-berlin.de

Abstract. Interoperability of digital libraries has been a research challenge for some time, especially its semantic aspect. In a digital library environment like Europeana, where a vast amount of information objects from heterogeneous sources will be made available, it is important to find standards that enable semantic interoperability. This dissertation aims at evaluating the applicability of CIDOC CRM for providing semantic interoperability in a digital library environment. To achieve this, the current use of CIDOC CRM in digital cultural heritage projects will be examined and its role for semantic interoperability will be determined. Limitations, benefits and prerequisites will influence the answer to the question on how CIDOC CRM can be integrated in Europeana in order to enhance semantic interoperability.

Keywords: Semantic interoperability, digital library, evaluation, CIDOC CRM, Europeana.

Significant Properties in the Preservation of Relational Databases

Ricardo André Pereira Freitas

CLEGI - Lusiada University, Vila Nova de Famalicão, Portugal
freitas@fam.ulusiada.pt

Abstract. Relational Databases are the most frequent type of databases used by organizations worldwide and are the base of several information systems. As in all digital objects, and concerning the digital preservation of them, the significant properties (significant characteristics) must be defined so that adopted strategies are appropriate. In previous work a neutral format (hardware and software independent) - DBML - was adopted to achieve a standard format used in the digital preservation of the relational databases data and structure. Currently, in this PhD project we walk further in the definition of the significant properties by considering the database semantics as an important characteristic that should also be preserved. For the representation of this higher level of abstraction we are going to use an ontology based approach. We will extract the entity-relationship model from the DBML representation and we will represent it as an ontology.

Keywords: Digital Preservation, Significant Properties, Significant Characteristics, Relational Databases, Ontology, OAIS, XML, Digital Objects.

Leveraging User Interaction and Collaboration for Improving Multilingual Information Access in Digital Libraries

Juliane Stiller

Berlin School of Library and Information Science, Humboldt University, Berlin, Germany
juliane.stiller@ibi.hu-berlin.de

Abstract. Evaluation of interactive cross-lingual information retrieval systems has been the focus of recent research. The goal is to support the users in formulating effective queries and selecting the documents which satisfy their information needs regardless of the language of the documents. This dissertation aims at harnessing the user-system interaction, extracting the added value and integrating it back into the system to improve the cross-lingual information retrieval system for successive users.

To achieve this, user input at different interaction points will be evaluated. This will among others include interaction during user-assisted query translations, implicit and explicit relevance feedback and social tags. To leverage this input, explorative studies need to be conducted to establish user input which might be beneficial and the methods to extract it. The dissertation wants to extend the scope of interactive cross-lingual information retrieval by harnessing user input as a mean for improving cross-lingual information retrieval tools.

Keywords: Digital Libraries, Interactive cross-language information retrieval, Social tags.

Workshop: Making Digital Libraries Interoperable: Challenges and Approaches

Donatella Castelli¹, Yannis Ioannidis², and Seamus Ross³

¹ Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy

² National and Kapodistrian University of Athens Panepistimiopolis, Greece

³ University of Toronto, Canada

The central theme of this Workshop is Digital Library Interoperability. Interoperability is a multi-layered and context-specific concept. It encompasses different levels along a multidimensional spectrum ranging from organisational to technological aspects. The Workshop addresses this challenging area from several perspectives: content, user, functionality, policy, quality, and architecture. Contributions will focus on relevant Digital Library interoperability aspects from conceptualisation at a high organisational level to instantiation at process level, as well as modelling techniques for representing and enabling interoperability between heterogeneous digital library mediation approaches, methods, and systems.

There are some scientific events which address the interoperability problem. However, they address only one dimension of the problem, usually, content interoperability or service interoperability. Our findings show the problem to be much more complex. The proposed ECDL2010 Workshop is aimed at addressing the dimensions of the Digital Library interoperability challenge much more broadly.

The workshop examines current approaches and new research directions for addressing the digital library interoperability challenge from a six faceted approach. The goal of this workshop is to provide researchers, practitioners and digital library developers with a forum fostering a constructive exchange of ideas on interoperability in digital libraries. The workshop at ECDL2009 raised many new questions related to interoperability and conceptualisation of knowledge, this purposed workshop offers an opportunity to take these discussions further and to reflect on both research conducted in other projects and taken forward by DL.org engaging many of the participants who took part in our 1st Workshop at ECDL2009.

Workshop: Networked Knowledge Organisation Systems and Services

Traugott Koch¹, Marianne Lykke Nielsen², and Douglas Tudhope³

¹ Max Planck Digital Library, Berlin, Germany

² Royal School of Library and Information Science, Denmark

³ University of Glamorgan, UK

The 9th NKOS workshop at ECDL explores the potential of Knowledge Organization Systems, such as classification systems, taxonomies, thesauri, ontologies, and lexical databases. These tools attempt to model the underlying semantic structure of a domain for purposes of information retrieval, knowledge discovery, language engineering, and the semantic web. The workshop provides an opportunity to report and discuss projects, research, and development related to Networked Knowledge Organization Systems/Services in next-generation digital libraries.

ECDL is the established venue for reporting on European NKOS activities, complementing the US series of workshops. The workshop allows major projects to report results, newcomers to interact with established people in the field and discussion of topical issues, requiring consensus or coordination, including standards efforts. Thus previous workshops have seen focused discussion on early drafts of BSI and ISO KOS standards, the W3C SKOS standard, the interface between traditional Library Science vocabularies and Semantic Web efforts, social tagging and its relation to established vocabularies, KOS metadata and the different types of KOS. The ECDL venue affords participation by KOS researchers and developers from different perspectives (reflecting the different conference threads), such as KOS design and construction, API and service developers, user oriented issues, management of KOS in registries.

Workshop: Very Large Digital Libraries

Yannis Ioannidis¹, Paolo Manghi², and Pasquale Pagano²

¹ National and Kapodistrian University of Athens Panepistimiopolis, Greece

² Istituto di Scienza e Tecnologie dell'Informazione, Pisa, Italy

The implementation of modern Digital Libraries is more demanding than in the past. Information consumers are facing with the need to access ever growing, heterogeneous, possibly federated Information Spaces while information providers are interested in satisfying such needs by sharing rich and organised views over their information deluge. Because of their fundamental role of information production and dissemination vehicle, Digital Libraries are also expected to provide information society with functionalities and services that must be available 24/7 and guarantee the expected quality of service. This scenario leads to the development of Very Large Digital Libraries, which are very large in terms of the number of information objects and collections to be made available, users to be served and potentially distributed functionality/content resources needed to construct them. Research on VLDLs opens up novel and actual scenarios, where researchers have to confront with new foundational and system design challenges in a context having scalability, interoperability and sustainability as focal points. Authors and participants of the past editions of Very Large Digital Library Workshop, respectively at ECDL 2008 and ECDL 2009, have confirmed the importance of the topic and eagerly started investigating the foundations of this new and hot research field (results have been published at SIGMOD Record and D-Lib Magazine). The goal of the Third Very Large Digital Library workshop is to prosecute such fertile discussions, hence to continue on providing researchers, practitioners and application developers with a forum fostering a constructive exchange among all key actors in the field of Very Large Digital Libraries.

Tutorial: Exploring Perspectives on the Evaluation of Digital Libraries

Giannis Tsakonas and Christos Papatheodorou

Ionian University, Corfu, Greece

During the last twenty years, the digital library domain has exhibited a significant growth aiming to fulfill the diverse information needs of heterogeneous user communities. Digital libraries, either existing as research prototype in a research center or laboratory, or operating in an intense environment enjoying actual usage from end users, have explicitly the need to measure and evaluate their operation. Digital library evaluation is a multifaceted domain aiming to compose the views and perspectives of various agents, such as digital library developers, librarians, curators, information and computer scientists. Several research fields, like information retrieval, human computer interaction, information seeking, user behavior analysis, organization and management of information systems, are contributing to capture, analyze and interpret data into useful suggestions of beneficial value for the information provider and its users.

This half-day tutorial will describe the current state of the art on digital libraries evaluation focusing to the following critical questions that project managers, digital library developers and librarians face: the motivations forcing to evaluate, how these motivations are connected to methodologies, techniques and criteria, how effective is one methodology compared to another in relation to the context of operation, what are the appropriate personnel and resources, as well as the organizational and legal requirements for conducting an evaluation experiment and what are the expected derivatives.

The tutorial is divided in two sections, each of them counting ninety minutes duration:

1. The first section is dedicated to outlining the digital library evaluation dimensions and approaches, the methodologies, criteria, metrics and measurement instruments employed in evaluation activities. Through the enumeration of important projects and initiatives, the pros and cons of each approach will be sketched.
2. The second part of the tutorial will start with a formal model presenting the dominant concepts of the evaluation research field, providing a conceptual synopsis of the issues discussed in the first part. Furthermore this section includes a practical session in which several indicative real-life tasks will be given to the participants and their response to the evaluation challenges will be discussed.

Tutorial: Introduction to (Teaching/Learning about) Digital Libraries

Edward Fox

Virginia Tech, Blacksburg, USA

This tutorial will provide a thorough and deep introduction to the DL field, introducing and building upon a firm theoretical foundation (starting with 5S: Streams, Structures, Spaces, Scenarios, Societies as well as the DELOS Reference Model), giving careful definitions and explanations of all the key parts of a minimal digital library, and expanding from that basis to cover key DL issues, illustrated with a well-chosen set of case studies. Results from an NSF grant to develop DL curriculum will be presented, including descriptions (aimed at CS or LIS teachers and learners) of the major modules and sub-modules that cover the core DL topics and related topics. There also will be a brief demonstration of digital preservation visualizations being taught at the Digital Preserve island in Second Life.

The tutorial has five main parts:

1. Theoretical Foundations: DELOS Reference Model and the ‘Ss: Societies, Scenarios, Spaces, Structures, Streams (covering content; multimedia; digital objects; metadata; ontologies; indexes; classification; retrieval models; user interfaces; information access; logging; DL communities (librarians, patrons, ...); IP; sustainability; functionality; policies;
2. Higher Level DL Constructs: Collections, Catalogs, Repositories, Handles, Interoperability, Standards, Scalability, Services taxonomy and services; Systems; Case Studies
3. Advanced Topics: Quality, Integration, How to Build a DL, Lessons from Ensemble
4. Digital Library Curricular Resources: Overview; Digital Objects; Collection Development; Info/Knowledge Organization; Architecture; User Behavior / Interactions; Services; Preservation; Management and Evaluation; DL Education and Research
5. Digital Preservation in Second Life: Visualizations for education

Tutorial: Memento and Open Annotation

Michael L. Nelson¹, Robert Sanderson², and Herbert Van de Sompel²

¹ Old Dominion University, Norfolk, USA

² Los Alamos National Laboratory, USA

This tutorial will introduce and provide technical details regarding two emerging frameworks that are of significant importance to both Web-based scholarly communication and the Web at large:

- Memento: The Memento framework essentially introduces the time component that has been missing from the Web. Memento allows to seamlessly HTTP-navigate from the URI of a resource to archived versions of that resource by adding a timestamp to HTTP GET requests. The result is the integration of the current and the past (archived) Web.
- Open Annotation: The Open Annotation Collaboration works towards defining and deploying a Web-centric interoperable annotation framework aimed at sharing annotations across the boundaries of annotation clients, content collections, and Web resources in general. In order to support deployment in the Web at large, the Open Annotation approach strictly adheres to the Architecture of the Web and to Linked Data principles. In addition, it takes into account requirements imposed by scholarly applications such as annotating multiple targets, and achieving robustness of annotations over time.

In addition, to illustrate the value added by the two emerging frameworks to Web-based scholarship, the tutorial will present the results of research that explored the combination of the temporal capabilities proposed by Open Annotation, and the time-travel capabilities offered by Memento, as an approach to realize an annotation framework that provides guarantees regarding the robustness of Web-annotations over time.

Tutorial: Multimedia Document Access

Stefan Ruger

Open University, Milton Keynes, UK

Computer technology has changed our access to information tremendously: We used to search authors or titles (which we had to know) in library cards in order to locate relevant books; now we can issue keyword searches within the full text of unimaginably large book repositories to identify the authors, titles and locations of relevant resources. What about the corresponding challenge of finding multimedia by fragments, examples and excerpts? Rather than asking for a music piece by artist and title, can we hum its tune to find it? Can doctors submit scans of a patient to identify medically similar images of diagnosed cases in a database? Can your mobile phone take a picture of a statue and return to you resources about its artist via a service that it sends this picture to?

Some of the challenges of these questions are given by the semantic gap between what computers can index and high-level human concepts; related to this is an inherent technological limitation of automated annotation of images from pixels alone. Other challenges are given by polysemy, ie, the many meanings and interpretations that are inherent in visual material and the corresponding wide range of a users information need.

This tutorial will demonstrate how these challenges can be tackled by automated processing and machine learning and by utilising the skills of the user, for example through harnessing and directing browsing activities with relevance feedback, thus putting the user centre stage. Other automated processing methods that discover and utilise world knowledge in the form of wikipedia (an online, linked, multilingual and open content encyclopedia) will be shown to not only improve multimedia retrieval but also give surprising insights into the human nature.

Panel: Developing Services to Support Research Data Management and Sharing

Liz Lyon¹, Joy Davidson², Veerle Van den Eynden³,
Robin Rice⁴, and Rob Grim⁵

¹ University of Bath, UK

² Digital Curation Centre, UK

³ University of Essex, UK

⁴ University of Edinburgh, UK

⁵ Tilburg University, Netherlands

Effective research data management (RDM) is gaining increasing importance as funders, publishers, and research institutions voice concerns about the loss of data associated with funded, published research and its lack of availability and accessibility beyond the life of the research project. The pressure on academics to manage, document, share and preserve their data is not balanced by incentives, support or mechanisms for them to do so.

Digital library systems are becoming increasingly sophisticated and the integration of a variety of forms of research outputs beyond traditional publications and including research data is in sight. Users are also becoming increasingly sophisticated and expect not only to read these outputs but to link, merge, analyse, visualise and manipulate them for their own purposes. Creative Commons licenses are helping to remove legal barriers to repurposing and mashing digital objects to create new forms of knowledge. Metadata standards are increasingly appearing for more and more disciplinary data types that can aid in the control of their storage and dissemination.

Currently there is a disconnect or gap between the readiness of the current information environment to deal with large amounts of heterogeneous data and the disposition of academics in research institutions to update their practices of data management and curation. Furthermore there is a well-documented reluctance on the part of researchers to deposit their data assets in trusted repositories for the benefit of unknown users.

The panel will focus on this gap between digital library systems and researchers current practice for managing data and explore appropriate services that can help to bridge it.

Some of the questions for discussion at the panel are: (1) Who is responsible for research data produced within public institutions? (a) when during the research project? (b) when after the research project? (for how long?) (2) Who are the users of research data that is shared and what are their needs? (3) What research data needs to be curated and what doesn't? (4) What kinds of tools and services can ease the burden on researchers? (5) How should institutions fulfil their responsibilities for managing research data? (6) How is heterogeneous data best managed and shared? (7) Why don't researchers manage or curate their data well? (8) Why should researchers share data openly?

Author Index

- Agerbæk Black, Esben 393
Agosti, Maristella 397
Aitken, Brian 401
Allen, Robert B. 46
Amato, Giuseppe 55
Andrade, Felipe 466
Andrews, Pierre 327
Angelis, Stavros 445
Artini, Michele 14
Athanasopoulos, George 405
Audenaert, Neal 307
Autayeu, Aliaksandr 327
- Baillie, Mark 196
Bainbridge, David 168
Balke, Wolf-Tilo 30
Bardi, Alessia 14
Batjargal, Biligsaikhan 518
Beckers, Thomas 409
Beel, Jöran 413, 449
Benz, Dominik 417
Bernard, Jürgen 352
Berndt, René 376
Bia, Alejandro 136, 421
Billerbeck, Bodo 273
Birrell, Duncan 104, 510
Biryukov, Maria 228
Blandford, Ann 184
Blanke, Tobias 522
Blümel, Ina 376
Bogen, Paul 116
Böhm, Klemens 156
Bolettieri, Paolo 55
Bountouri, Lina 38
Brase, Jan 352
Briston, Heather 437
Brusilovsky, Peter 116, 506
Bryant, Michael 522
Buchanan, George 92, 433
Buchanan, Steven 425
Buchmann, Erik 156
- Cabanac, Guillaume 340
Caicedo, Juan C. 429
Calabretto, Sylvie 364
- Camargo, Jorge E. 429
Candela, Leonardo 14
Canós, José H. 453, 526
Carpenter II, B. Stephen 506
Cassel, Lillian 116, 506
Castillo, Carlos J. 526
Cerviño Beresi, Ulises 196
Chang, Yung-Ting 389
Chanod, Jean-Pierre 538
Chen, Shu-Jiun 389
Cisco, Davide 397
Clausen, Michael 376
Constantopoulos, Panos 445
Cornelis, Chris 457
Crestani, Fabio 474
- Damm, David 376
da S. Torres, Ricardo 466, 486
Déjean, Hervé 538
Delcambre, Lois 116, 506
Demartini, Gianluca 273
Diet, Jürgen 376
Di Nunzio, Giorgio Maria 397
Diriye, Abdigani 184
Dobрева, Milena 510
Dodd, Helen 433
Dong, Cailing 228
Donnelly, Martin 530
Dumais, Susan 1
Dunsire, Gordon 104
- Edwards, Stephen 506
Eichler, Kathrin 534
Estlund, Karen 437
- Fachry, Khairun Nisa 409
Feliciati, Pierluigi 510
Fellner, Dieter 352, 376
Firan, Claudiu S. 273
Fox, Edward 116, 405, 466, 506,
514, 550
Freire, Nuno 441
Fremerey, Christian 376
Friedrich, Nick 413

- Fuhr, Norbert 409
 Furuta, Richard 22, 116, 307, 506

 Gaitanou, Panorea 38
 Garcia, Daniel D. 116, 506
 Gavrilis, Dimitris 445
 Gennaro, Claudio 55
 Gergatsoulis, Manolis 38
 Gerlach Sanches Lima, Angela 417
 Gipp, Bela 413, 449
 Giunchiglia, Fausto 327
 Gómez, Jaime 136, 421
 González, Fabio A. 429
 Granitzer, Michael 315
 Greifeneder, Michael 124

 Hagen, Matthias 384
 Halle, Axel 417
 Hallerman, Eric 466
 Haughton, Tim 294
 Hedges, Mark 522
 Heidinger, Clemens 156
 Hemmje, Matthias 470
 Hensen, Holmer 534
 Herring, Matthew 502
 Hislop, Gregory 506
 Hoare, Cathal 208
 Hoffmann, Oliver 216
 Hotho, Andreas 417
 Huber, Matthias 156
 Hubert, Gilles 340
 Hurtado Martín, Germán 457

 Iatropoulou, Katerina 294
 Ioannidis, Yannis 405, 542
 Iofciu, Tereza 273

 Jackson, Andrew 401
 Jacquin, Thierry 538
 Jäschke, Robert 417
 Jones, Matt 433
 Jones, Sarah 530

 Kakalettris, George 405
 Kamps, Jaap 248
 Kanan, Tarek 514
 Kanhabua, Nattiya 261
 Kapidakis, Sarantos 546
 Karadkar, Unmil 22
 Kazai, Gabriella 294

 Kern, Roman 315, 461
 Kimura, Fuminori 518
 Kim, Yunhyong 196
 Klakow, Dietrich 482
 Klas, Claus-Peter 470
 Klein, Reinhard 376
 Koepler, Oliver 352
 Kohlhammer, Jörn 352
 Konstantelos, Leo 148
 Korner, Christian 461
 Kozievitch, Nádia P. 466
 Krahl, Frank 376
 Krestel, Ralf 273
 Kyriakidi, Marialena 542

 Landwich, Paul 470
 Lease, Matthew 482
 Lee, Spencer 550
 Le Meur, Jean-Yves 236
 Lempesis, Antonis 294
 Lindley, Andrew 401
 Llavador, Manuel 453, 526
 Lucchese, George 307

 Maeda, Akira 518
 Manghi, Paolo 14, 294
 Manola, Natalia 294, 405
 Marante, María Isabel 453
 Marian, Ludmila 236
 Markov, Ilya 474
 Masiero, Ivano 397
 Maurizio, Marek 478
 McCulloch, Emma 510
 McHugh, Andrew 148
 McMenemy, David 425
 Meghini, Carlo 2, 405
 Meintanis, Konstantinos 80
 Melucci, Massimo 397
 Menzies, Kathleen 104
 Michaeler, Edith 401
 Middleton, Bo 502
 Mikulicic, Marko 14, 294
 Milic-Frayling, Natasa 294
 Momtazi, Saeedeh 482
 Müller-Quade, Jörn 156
 Muñoz, Pablo 526
 Muñoz, Rafael 136, 421
 Murthy, Uma 466

 Naessens, Helga 457
 Neumann, Günter 534

- Nordlie, Ragnar 409
 Nørvåg, Kjetil 261
 Novak, Brook J. 168

 Orsini, Renzo 478

 Palacio, Damien 340
 Palmer, Richard 522
 Papatheodorou, Christos 38, 445
 Park, Yung Ah 22
 Pattenden-Fail, John W. 530
 Pearson, Jennifer 92
 Penatti, Otávio A.B. 486
 Peng, Pei-Wen 389
 Petrov, Petar 124
 Pharo, Nils 409
 Pirkola, Ari 490
 Portier, Pierre-Édouard 364
 Potter, Ned 502

 Rajman, Martin 236
 Rauber, Andreas 124, 405
 Rechert, Klaus 494
 Reithinger, Norbert 534
 Reitz, Florian 216
 Richardson, Gramm 498
 Ross, Seamus 401
 Ruppert, Tobias 352
 Ruthven, Ian 196, 510
 Ruzzoli, Felix 494

 Sallaberry, Christian 340
 Savino, Pasquale 55
 Scherer, Maximilian 376
 Schmeier, Sven 534
 Schockaert, Steven 457
 Schreck, Tobias 352, 376
 Schumacher, Kinga 534
 Seifert, Inessa 534
 Sens, Irina 352, 376
 Sfakakis, Michalis 546
 Shaker, Ammar 413
 Shipman, Beccy 502
 Shipman, Frank M. 80, 116, 506
 Soergel, Dagobert 405
 Song, Dawei 196

 Sorensen, Humphrey 208
 Spyratos, Nicolas 2
 Staikos, Panagiotis 546
 Stamatogiannakis, Lefteris 542
 Steenweg, Helge 417
 Stefani, Sven 417
 Stein, Benno Maria 384
 Strodl, Stephan 124
 Strohmaier, Markus 461
 Stumme, Gerd 417
 Sugibuchi, Tsuyoshi 2
 Suleman, Hussein 550
 Sykes, Jonathan 510

 Talvensaaari, Tuomas 490
 Taylor, Adriana 449
 Thomas, Verena 376
 Tombros, Anastasios 184
 Toms, Elaine G. 282
 Tönnies, Sascha 30
 Toze, Sandra 282
 Triantafyllidi, Mei Li 542

 Ünal, Yurdagül 510

 Valero, Héctor 526
 Valizada, Isgandar 494
 van den Dobbelsteen, Maurice 401
 Vayanou, Maria 542
 Versfeld, Rizmari 550
 Vesely, Martin 236
 von Suchodoletz, Dirk 494

 Wartena, Christian 176
 Welte, Randolph 494
 Wessel, Raoul 376
 Westman, Stina 67
 Williams, Kyle 550
 Wu, Diane 389

 Yang, Seungwon 514
 Yudelson, Michael 116

 Zarro, Michael A. 46
 Zhang, Junte 248