Stefan Gradmann
Francesca Borri
Carlo Meghini
Heiko Schuldt (Eds.)

LNCS 6966

Research and Advanced Technology for Digital Libraries

International Conference on Theory and Practice of Digital Libraries, TPDL 2011
Berlin, Germany, September 2011, Proceedings

Springer

Lecture Notes in Computer Science 6966

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Stefan Gradmann Francesca Borri
Carlo Meghini Heiko Schuldt (Eds.)

Research and Advanced Technology for Digital Libraries

International Conference on Theory and Practice of Digital Libraries, TPDL 2011
Berlin, Germany, September 26-28, 2011
Proceedings

Springer

Volume Editors

Stefan Gradmann
Humboldt-Universität zu Berlin
Berlin School of Library and Information Science, Berlin, Germany
E-mail: stefan.gradmann@ibi.hu-berlin.de

Francesca Borri
Carlo Meghini
Istituto di Scienza e Tecnologie dell'Informazione
Consiglio Nazionale delle Ricerche, Pisa, Italy
E-mail:{francesca.borri, carlo.meghini}@isti.cnr.it

Heiko Schuldt
University of Basel
Databases and Information Systems Group, Basel, Switzerland
E-mail: heiko.schuldt@unibas.ch

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-24468-1 e-ISBN 978-3-642-24469-8
DOI 10.1007/978-3-642-24469-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011937029

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

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

We are happy to present the proceedings of the 15th edition of the TPDL 2011 conference, which is part of an impressive series starting with the first European Conference on Research and Advanced Technology for Digital Libraries (ECDL) in Pisa (1997) and subsequently Heraklion (1998), Paris (1999), Lisbon (2000), Darmstadt (2001), Rome (2002), Trondheim (2003), Bath (2004), Vienna (2005), Alicante (2006), Budapest (2007), Aarhus (2008), Corfu (2009) and Glasgow (2010). In the course of these years, ECDL had become one of the major international reference meetings for an ever-growing and more and more multidisciplinary community and thus has considerably broadened in geographical scope: what started as a European event has grown into a part of the increasingly integrated international community building around the notion of digital libraries. From this perspective it was a logical step to rename the conference: Humboldt University in Berlin was thus proud to host the first conference named TPDL, standing for the Conference on Theory and Practice of Digital Libraries. The necessity to avoid acronym conflicts with the European Computer Driving Licence thus coincided with the need to see the international scope of the conference reflected in a new name that would avoid the 'regional' limitation to just one continent.

But the conference has not only broadened in geographical scope: it has also diversified thematically! What started as mostly a computer science-driven event has now grown into a truly interdisciplinary conference—it is therefore logical that the conference was organized by the Berlin School of Library and Information Science!

Thematically, TPDL 2011 was built along two main strands: Technology and Methodology as a foundational layer as well as Applications and User Experience as a practical layer. The conference thus started with three ambitious tutorials about "Harmonizing Models for the Digital World", on "Distributed Cloud-Based Collaboration" and on "Aggregation and Reuse of Digital Objects Metadata from Distributed Digital Libraries", with a Doctoral Consortium in parallel pre-senting work by seven students who were given feedback by experts in digital library research. The technical program included sessions on Networked Information, Semantics and Interoperability (twice!), Systems and Architectures, Text and Multimedia Retrieval, Collaborative Information Spaces, DL Applications and Legal Aspects, User Interaction and Information Visualization, User Studies, Archives and Repositories, Europeana and Preservation. In addition, we had keynote speeches by eminent colleagues such as Moira C. Norrie, Clifford Lynch and Thomas Hofmann. Following the main conference, TPDL 2011 hosted five parallel one-day workshops, among which were "old acquaintances" such as The 10th European Networked Knowledge Organization Systems (NKOS) Workshop or the 4th Workshop on Very Large Digital Libraries (VLDL 2011),

but also new ones such as the ones on Semantic Digital Archives—sustainable long-term curation perspectives of Cultural Heritage, on Linking Research and Education in Digital Libraries or again on User-Oriented Evaluation of Digital Library Interfaces.

For TPDL 2011, we set up two different Program Committees—one for full and short research papers, and one for demos. Furthermore, we did not launch a dedicated call for posters but invited all authors of accepted full, short, and demo papers to also present their work as a poster in order to promote the poster sessions and at the same time increase the visibility of all research papers. For the review process, we adopted the successful two-tier model from ECDL 2010: a first-tier Program Committee of 69 members was supervised by 17 senior Program Committee members. Each paper was viewed independently by three members of the Program Committee; for each paper, the subsequent discussion and disambiguation of review results was moderated by one member of the senior Program Committee.

TPDL 2011 attracted in total 162 submissions from all over the world (141 full and short papers, 21 demo papers) out of which the two committees 27 full papers (19%), 13 short papers (9%), and 9 demos (43%). Moreover, nine additional submissions were accepted as posters (6%) to be presented during the TPDL poster session.

The success of TPDL2011 in all respects would not have been possible without the tremendous contributions of all members of the Organizing Committee and the Program Committee as well as the numerous student volunteers that supported the conference in its various stages. Special thanks go to the Program Chairs, Carlo Meghini and Heiko Schuldt, as well as to Marlies Olensky as the Local Organizing Chair for their outstanding and continuous work. Finally, we would like to thank the sponsoring organizations Emerald Group Publishing, Ex Libris, Swets Information Services, IOS Press, Ashgate Publishing Group and The Coalition for Networked Information (CNI) for their substantial support in spite of the economic challenges of our times.

October 2011

Stefan Gradmann
Francesca Borri
Carlo Meghini
Heiko Schuldt

Organization

TPDL 2011 was organized by the Institut für Bibliotheks- und Informationswissenschaft, Philosophische Fakultät I of Humboldt-Universität zu Berlin, Germany

Organizing Committee

General Chair	Stefan Gradmann, Humboldt-Universität zu Berlin, Germany
Local Chairs	Michael Seadle, Humboldt-Universität zu Berlin, Germany
	Peter Schirmbacher, Humboldt-Universität zu Berlin, Germany
	Wolfgang Coy, Humboldt-Universität zu Berlin, Germany
Local Organizing Chair	Marlies Olensky, Humboldt-Universität zu Berlin, Germany
Local Organizing Committee	Peggy Beßler, Humboldt-Universität zu Berlin, Germany
	Jenny Sieber, Humboldt-Universität zu Berlin, Germany
	Juliane Stiller, Humboldt-Universität zu Berlin, Germany
Proceedings Chairs	Costantino Thanos, ISTI-CNR, Italy
	Francesca Borri, ISTI-CNR, Italy
Publicity Chairs	Laszlo Kovacs, MTA SZTAKI, Hungary
	Axel Kaschte, Ex Libris, Germany

Program Committee

Program Chairs	Carlo Meghini, ISTI-CNR, Italy
	Heiko Schuldt, University of Basel, Switzerland
Poster Chair	Susanne Dobratz, Humboldt University Library, Germany
Demo Chair	Thomas Risse, L3S Research Center, Germany
Panel Chair	Johan Oomen, Netherlands Institute for Sound and Vision, The Netherlands
Tutorial Chair	Vivien Petras, Humboldt-Universität zu Berlin, Germany
Workshop Chair	Nicola Ferro, University of Padova, Italy
Doctoral Consortium Chair	Milena Dobreva, University of Strathclyde, UK
Best Paper Award Chair	Maristella Agosti, University of Padova, Italy

Meta-Reviewers

Maristella Agosti	University of Padova, Italy
Margherita Antona	Foundation for Research and Technology (FORTH), Greece
Tobias Blanke	King's College London, UK
Lou Burnard	TGE-Adonis, UK
Donatella Castelli	ISTI-CNR, Italy
Edward Fox	Virginia Tech, USA
Max Kaiser	Austrian National Library, Austria
Stephanos Kollias	National Technical University of Athens, Greece
Erwin Laure	KTH, Sweden
Andreas Rauber	Vienna University of Technology, Austria
Laurent Romary	LORIA, France
Fabrizio Sebastiani	ISTI-CNR, Italy
Ingeborg T. Solvberg	Norwegian University of Science and Technology, Norway
Hussein Suleman	University of Cape Town, South Africa
Manfred Thaller	University of Cologne, Germany
Herbert van de Sompel	Los Alamos National Laboratory, USA
Peter Wittenburg	Max Planck Institute for Psycholinguistics, The Netherlands

Program Committee

Sheila Anderson	King's College London, UK
David Bainbridge	University of Waikato, New Zealand
Christian Bizer	FU Berlin, Germany
Jose Borbinha	IST/INESC-ID - Information Systems Group, Portugal
Jan Brase	TIB Hannover, Germany
Pavel Braslavski	Ural State University, Russia
George Buchanan	City University London, UK
B. Barla Cambazoglu	Yahoo! Research, Spain
Stavros Christodoulakis	Technical University of Crete, Greece
Greg Crane	Tufts University, USA
Fabio Crestani	University of Lugano, Switzerland
Sally Jo Cunningham	University of Waikato, New Zealand
Theodore Dalamagas	National Technical University of Athens, Greece
Alberto del Bimbo	University of Florence, Italy
J. Stephen Downie	University of Illinois, USA
Schubert Foo	Nanyang Technological University, Singapore
Nuno Freire	The European Library, Portugal
Johann-Christoph Freytag	Humboldt University Berlin, Germany

Ingo Frommholz	University of Glasgow, UK
Norbert Fuhr	University of Duisburg-Essen, Germany
Richard Furuta	Texas A&M University, USA
Marcos André Goncalves	Federal University of Minas Gerais, Brazil
Julio Gonzalo	Universidad Autonoma de Madrid, Spain
Jane Greenberg	University of North Carolina at Chapel Hill, USA
Maria Guercio	University of Urbino, Italy
Preben Hansen	Swedish Institute of Computer Science, Sweden
Mark Hedges	King's College London, UK
Geneva Henry	Rice University, USA
Wolfram Horstmann	University Library Bielefeld, Germany
Jieh Hsiang	National Taiwan University, Taiwan
Antoine Isaac	Vrije Universiteit Amsterdam, The Netherlands
Sarantos Kapidakis	Ionian University, Greece
Claus-Peter Klas	FernUniversität Hagen, Germany
Wolfgang Klas	University of Vienna, Austria
Traugott Koch	Max Planck Digital Library, Germany
Marc Kuester	University of Applied Sciences Worms, Germany
Carl Lagoze	Cornell University, USA
Audrey Laplante	Université de Montral, Canada
Gerhard Lauer	University of Göttingen, Germany
Annelise Mark Pejtersen	Riso National Laboratory, Denmark
Andras Micsik	SZTAKI, Hungary
Diego Milano	University of Basel, Switzerland
Reagan Moore	San Diego Supercomputing Center, USA
John Mylopoulos	University of Trento, Italy
Wolfgang Nejdl	L3S and University of Hannover, Germany
Erich Neuhold	University of Vienna, Austria
Claudia Niedere	L3S, Germany
Kjetil Norvag	Norwegian University of Science and Technology, Norway
Pasquale Pagano	ISTI-CNR, Italy
Christos	Papatheodorou, Ionian University, Greece
Nils Pharo	Oslo University College, Norway
Jean-Luc Pinol	TGE ADONIS, France
Harald Reiterer	University of Konstanz, Germany
Seamus Ross	University of Toronto, Canada
Robert Sanderson	Los Alamos National Laboratory, USA
Felix Sasaki	DFKI, Germany
Rudi Schmiede	Darmstadt University of Technology, Germany
Susan Schreibman	Digital Humanities Observatory, Ireland
Ray Siemens	University of Victoria, Canada
Beat Signer	Vrije Universiteit Brussel, Belgium

Fabio Simeoni	University of Strathclyde, UK
Nicolas Spyratos	Université de Paris-Sud, France
Ulrike Steffens	OFFIS, Germany
Shigeo Sugimoto	University of Tsukuba, Japan
Elaine Toms	Dalhousie University, Canada
Doug Tudhope	University of Glamorgan, UK
Can Türker	Functional Genomics Center Zürich, Switzerland
Felisa Verdejo	National Distance Learning University (UNED) Madrid, Spain
Christian Wolff	University of Regensburg, Germany

Program Committee for Demos

Wolf-Tilo Balke	TU Braunschweig, Germany
Bernhard Haslhofer	Cornell University, USA
Wim Peters	University of Sheffield, UK
Dimitri Skoutas	Technical University of Crete, Greece
Marc Spaniol	Max Planck Institute for Informatics, Germany
Xuan Zhou	Renmin University of China, China
Ismail Sengor Altingövde	L3S Research Center, Germany
Stefan Dietze	Open University, UK
Dimitrios Tsoumakos	National Technical University of Athens, Greece

Sponsors

Emerald Group Publishing
Ex Libris
Swets Information Services
IOS Press
Ashgate Publishing Group
The Coalition for Networked Information (CNI)

Acknowledgements

The Organizing Committee and the Program Committee of TPDL 2011 sincerely thank Giuditta Feo of CNR ISTI for her support in the organization of the conference and in the preparation of the present conference proceedings.

Table of Contents

Keynotes

Technical Sessions

Networked Information

Semantics and Interoperability I

Systems and Architectures

Text and Multimedia Retrieval

Collaborative Information Spaces

DL Applications and Legal Aspects

User Interaction and Information Visualization

Semantics and Interoperability II

User Studies

Archives and Repositories

Europeana

Preservation

Demo Sessions

Poster Session

Panel Session

Doctoral Consortium

Paper, Pen and Touch

Moira C. Norrie

Institute for Information Systems, ETH Zurich
CH-8092 Zurich, Switzerland
norrie@inf.ethz.ch

Abstract. It has long been recognised by researchers that the affordances of paper are likely to ensure that it will continue to be in widespread use in the work place, homes and public spaces. Consequently, numerous research projects have investigated ways of integrating paper with digital media and services. In recent years, a lot of this research has revolved around the digital pen and paper technology developed by the Swedish company Anoto, since it offers a robust solution for tracking the position of a pen on paper. While the commercial sector has tended to focus on applications related to the capture of handwriting, many of these research projects have investigated the use of the pen for real-time interaction and possibilities of turning paper into an interactive medium.

Researchers were also quick to realise that digital pen and paper technology could be adapted to support other forms of pen-based interaction and have developed digital whiteboards and tabletops based on the technology. In addition, some systems have combined the technology with touch devices to support bimanual pen and touch interfaces. In the case of document manipulation, this means that touch could be used to perform actions such as a moving a document or turning pages, while the pen could be used to select elements within a document or to annotate it. Further, there are projects which have integrated the work on interactive paper and pen-based interaction on digital tabletops, investigating ways of allowing users to transfer document elements back and forth between paper and digital surfaces.

Despite the success of these research projects in terms of demonstrating the capabilities of digital pen and paper technology and how it could be exploited to support a wide variety of everyday tasks, there are still some technical and non-technical issues that need to addressed if there are to be major breakthroughs in terms of widespread adoption. The first part of the talk will review research in the field, while the second part will examine these issues and the way ahead.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, p. 1, 2011.

The Futures of Digital Libraries: The Evolution of an Idea

Clifford Lynch

Coalition for Networked Information

Abstract. The construction of digital libraries have certainly framed technological challenges, particularly with regard to various aspects of scale, and with the complexities of dealing with human languages, and indeed have given rise to substantial progress in these and other technical fields. But I believe that the greatest significance of digital libraries has been at a more profound intellectual level, inviting us to envision new kinds of environments for knowledge discovery, formulation, and dissemination; approaches to defining, managing and interacting with the cultural and intellectual record of our societies. We have repeatedly been forced to revisit questions of what constitutes a digital library, and how (indeed, even if) this differs from simply a collection of digitized or born-digital materials.

In my presentation I will look at some of there recent responses of these challenges of vision, examining emerging systems like Europeana and proposals like the Digital Public Library of America, a few well established operational digital libraries in various sectors, developing scientific knowledge management environments that integrate scholarship, scholarly communication and evidence, and even the changing ways in which we think of the collective mass of information available worldwide through the internet. As part of my analysis, I will discuss the enormous distorting effects of current copyright laws on our ability to realize many of our collective visions and to achieve necessary scale, but also the promise offered by the new renaissance in public engagement (citizen science, social media and related developments) and the progress of various movements towards openness. Finally, as I look to the conceptual future of digital libraries I'll consider the steady advance of technological capabilities in the analysis and exploitation of large, dynamic corpora of materials; these will also continue to reshape our understanding of future directions and possibilities.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, p. 2, 2011.

Connecting Archival Collections: The Social Networks and Archival Context Project

Ray R. Larson and Krishna Janakiraman

School of Information
University of California, Berkeley
Berkeley, California, USA, 94720-4600
{ray,krisha}@ischool.berkeley.edu

Abstract. This paper describes the Social Networks and Archival Context project, built on a database of merged Encoded Archival Context - Corporate Bodies, Persons, and Families (EAC-CPF) records derived from Encoded Archival Description (EAD) records held by the Library of Congress, the California Digital Library, the Northwest Digital Archives, and Virginia Heritage, combined with information from name authority files from the Library of Congress (Library of Congress Name Authority File), OCLC Research (The Virtual International Authority File), and the Getty Vocabulary Program (Union List of Artist Names). The database merges information from each instance of an individual name found in the EAD resources, along with variant names, biographical notes and their topical descriptions. The SNAC prototype interface makes this information searchable and browseable while retaining links to the various data sources.

1 Introduction

One of the important tasks of scholars is to use secondary (e.g. books and journal articles) and primary (original manuscripts, letters, etc.) research information in examining the lives, work, and events surrounding historic persons. For historians and many other scholars, the prefered sources are primary – the actual works or documents of the persons or organizations concerned, and to a lesser extent the interpretations of other scholars about those persons or organizations. In digital library research the focus has been largely on these secondary resources, and not on the original primary resources. In part this has been due to the relative lack of available metadata and digitized content for primary resources when compared to those of secondary resources.

The Social Networks and Archival Context (SNAC) project is trying to address this challenge of improving access to primary humanities resources through the use of advanced technologies. The project is producing software and developing open linked data resources that will enable scholars to connect historic persons to existing archival descriptions and to library catalogs and authority files. Thus creating a powerful new resource that enhances access to and understanding of the cultural resources in our archives, libraries, and museums

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 3–14, 2011.

through the description of the people who created them and whose lives are reflected in those resources. This new resource can serve a wide variety of objectives to benefit scholars, educators, students, and anyone interested in the record of our past. The goals of the SNAC project are to:

1. Support scholars and other users in discovering and identifying persons, families, and organizations, by making the names used by and for them searchable.
2. To merge together information from a wide variety of different sources by and about people and organizations and thus to enhance access to primary and secondary resources.
3. To discover and provide access to the social and professional networks within which people lived and worked by systematically documenting their relationships with one another to better understand the social-historical contexts within which the resources were created.
4. Provide archives, libraries and scholars with access to records describing persons, families, and organizations, thereby improving description of archival records and creating efficiencies in the re-use of metadata across repositories, and through open linked data resources (Figures 1, 2 and 3 show screens from our prototype interface).
5. Connect traditional library and archival information on persons, families and organizations with semantic web resources on the same persons, families and organizations, providing a resource for validation and contextual matching of traditional and semantic web resources.
6. Make available the software developed for matching persons, families and organizations based on the approaches developed during the project.

At the core of the SNAC project are the Encoded Archival Context - Corporate bodies, Persons, and Families (EAC-CPF) and the Encoded Archival Description (EAD) XML markup standards. SNAC is developing open-source software that will facilitate efficient and accurate derivation of authority control records from existing EAD archival finding aids from the Library of Congress (LoC) and three consortia, the Online Archive of California (OAC), the Northwest Digital Archive (NWDA), and Virginia Heritage (VH), and enhancing them with additional information in matching LoC, Getty Vocabulary Program, and Virtual International Authority File (VIAF) name authority records.

The SNAC project is being led by the Institute for Advanced Technology in the Humanities (IATH) at the University of Virginia, in collaboration with the California Digital Library (CDL), and the University of California, Berkeley School of Information (SI). The SNAC project is intended to benefit the humanities community most broadly, as it will assist the work of both archivists and users.

In the remainder of this paper we will describe the processes involved in creating the SNAC database and how it is presented through the public interface. The next section describes the extraction of EAC-CPF records from EAD records. We then examine how the names of individuals, organizations and families are

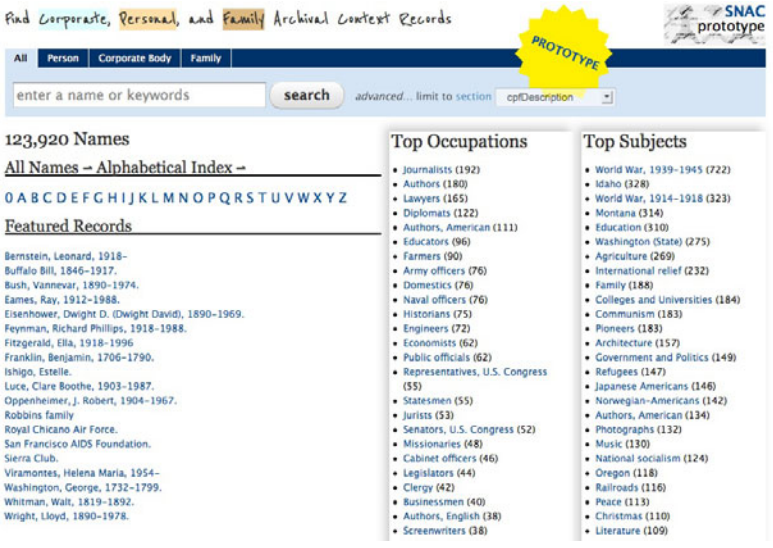


Fig. 1. Browsing Screen for SNAC Prototype

matched across the different sources of EAD data, and how they are matched with the library authority records from the Library of Congress and the Virtual International Authority File. In addition we present the results and effectiveness of different matching methods. Finally we discuss how the records are indexed and presented in the prototype public interface.

2 Extracting EAC from EAD

As mentioned above, the basic EAC-CPF records used in the SNAC project are derived from EAD archival descriptions. The current approach to doing this extraction uses the Extensible Markup-Language Transformation (XSLT 2.0) language in conjunction with the XPath 2.0 standard. Through the use of regular expressions and specially designed functions, the XSLT transform identifies elements of the EAD that represent individual persons, corporate bodies and families in various parts of the EAD record. Currently we have focused on the identification and extraction of individual records from the following EAD tag components: <persname>, <corpname>, and <famname> that occur within <origination>, <controlaccess>, and <unittitle>.

In the EAD records that we are using, the contributing archives have followed the "best practices" for encoding the <origination> and <controlaccess> elements, so these are usually formulated following strict cataloging rules (AACR2, for American archives and libraries). Most of our difficulties (in matching names and extracting contents) are caused by names that have not been formulated according to such rules. In some cases, for example, names are presented in direct order (John Smith) rather than inverted order (Smith, John), and may also be

Fig. 2. Example Record (George S. Patton)

combined with data that is not part of the name, (including subject subdivisions or uniform titles).

For each unique name string extracted by the above process, an EAC-CPF record is created. For records derived for creators (i.e., the source of the archival records), additional descriptive data for dates of existence, occupation, subject headings assigned to records, languages used, and biographical-historical information is extracted into the corresponding EAC-CPF records.

Once the EAD records have been processed, the result is a set of EAC-CPF records each containing a single identified name along with identification of the source EAD, and, in the case of creator records, any biographical information, dates of existence, etc. from the EAD source. Since EAC-CPF records are derived independently from each EAD record, there can be multiple records representing the same entity in multiple EAD collections. A key problem, then, is to identify multiple EAC-CPF records that represent the same entity and merge them together into a single record.

3 Matching Names in EAC-CPF and Authority Files

The next stage in processing the extracted EAC-CPF records attempts to combine the information of the individual records for the same entity into single records, and to enhance them with information derived from library authority files. This process involves matching the names in the EAC-CPF records with each other, and with the records in library authority files.

We currently have 158,079 EAC-CPF records: 114,639 persons, 41,177 corporate bodies and 2263 family names, derived from Library of Congress, the

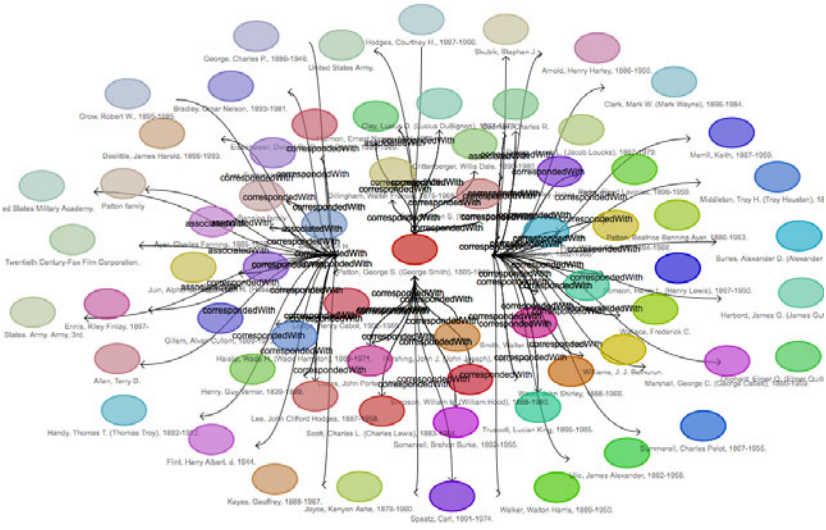


Fig. 3. Social Network (George S. Patton)

Online Archive of California and North West Digital Archive EAD records. The records were parsed with the EAC-CPF specification to extract information on name, type and relations of each entity. Preferred and alternate names from the VIAF name authority files were indexed using the Cheshire II information retrieval engine [3], which uses a probabilistic information retrieval algorithm to find matching VIAF records and their associated names given an entity name.

The simplest matching case occurs when the extracted names are already encoded using standard cataloging rules, and include existence dates. In this case, the extracted names will usually be identical to the names used by other EAD records that conform to the same practice and also with library authority records for the same person. This simple exact matching reduces the number of unique records to 129,915 from 158,079. We then assume that if the name being matched occurs in one of the library name authority records as a variant form, then the name is also a match. This was implemented using Cheshire and reduces the number of unique EAC-CPF records to 124,657. However, when there are variations in the name order, and variant forms of a name may be used in different EAD source records, then the matching problem becomes much more challenging, and warranted further investigation.

Our problem is similar to the well-studied named entity disambiguation problem, where the task is to identify the correct entity, in a given context, from a set of similar entities. Standard approaches use statistical learning techniques, either performing supervised learning and train classifiers that predict the relevance of an entity given a context or performing unsupervised learning and design clustering techniques that cluster similar entities together. As an example of the former, Bunescu and Pasca [2] suggest a method that trains Support Vector Machines (SVM) classifiers to disambiguate entities using the Wikipedia corpus.

The classifier was trained using features extracted from the title, hyperlinks linking other entities, categories assigned to the entity and Wikipedias redirect and disambiguate pages. Bagga and Baldwin [1], and Mann and Yarowsky [4], are examples of the latter technique, where similar entities are clustered using features extracted from entitys biographical information, words from sentences surrounding the entity in texts and entitys social network and relationships. Other techniques involve using gazetteers and name authority files as external references to aid the disambiguation process. Smith and Crane [5], for example, use gazetteers to disambiguate geographic place names.

3.1 Experiments in Name Disambiguation

Given a collection of entity/name pairs and a query name, the name disambiguation problem is to predict the entity that the query name should be associated with. Motivated by SNAC's access to the extensive VIAF collection, we experimented a supervised learning approach, based on the Naive Bayes classifier as an approach to solving this problem. Given an exhaustive collection of entities with a set of possible names for each of these entities, we compute simple representation of these names using shingles (also called character ngrams) and train naive bayes classifiers that predict the entity a query name belongs to using these representations.

We represent each name as a list of l length shingles. The shingles are computed by using a sliding window that is l characters in length. For example, the $l = 3$ shingles for the name Albert Einstein would be alb, lbe, ber, ert rt_ ,t_e, ... ein (where "_" represents white space between words).

This representation allows us to consider a name as a vector in a k dimensional vector space defined by the shingles, where k is the number of l length shingles in our collection. In addition to abstracting the structure of a name, this representation also allows us to create a shingle - entity reverse index. As would be explained in later sections, this index allows us to significantly reduce the space of candidate entities when a query is made.

3.2 Approaches

String Edit Distance. As a baseline approach, we experimented using string edit distance to rank entities based on a query name. Given a query name, we rank entities based on the mean edit distance between the names associated with the entity and the query name. We used the Levenshtein Distance to compute the edit distance between two strings.

Computing the edit distance over the entire space of entities is not feasible. To reduce the space, we only consider those entities that are indexed by the shingles present in the query name. Specifically, we first rank our entities based on how many shingles in the query name are indexed by them. We then pick top-M of these entities and score them based on the mean edit distance between the names associated with the entity and the query name. Entities with lower mean edit distance are ranked higher.

Naive Bayes. If we consider each entity as a class and names associated with the entity as examples of the class, we can cast the name disambiguation problem as a supervised statistical learning problem. Specifically, given a query name x_q, our problem then is to find an entity class C that is mostly likely to have generated the query name.

$$C = \arg\max_i P(C = c_i | X = x_q) \tag{1}$$

Using the Bayes rule, we can estimate the most likely C as follows,

$$C = \arg\max_i \frac{P(C = c_i)P(X = x_q | C = c_i)}{P(X = x_{q)}} \tag{2}$$

As $P(X = x_q)$ will be the same for all the entity classes, we can eliminate it from the above equation,

$$C \leftarrow \arg\max_i P(C = c_i)P(X = x_q | C = c_i) \tag{3}$$

The prior probability $P(C = c_i)$ can be estimated as,

$$P(C = c_i) = \frac{\#\{C = c_i\}}{D} \tag{4}$$

where D is the total number of name instances in the training collection and $\#\{C = c_i\}$ represents the number of name instances with entity class as c_i.

Based on our representation of names using shingles, we can consider each name as an instance of random vector in a k-dimensional space of shingles. The posterior probability in the above equation can then be written as,

$$P(X = x_q | C = c_i) = P(X = (x_{q1}, x_{q2}, x_{q3,}..., x_{qk}) | C = c_i) \tag{5}$$

where x_{qj} represents the event that a particular shingle j occurs in the query name. Estimating the above posterior probability would amount to estimating the joint distribution of the random vector X , assuming each component of the k -dimensional random vector to be binary, this would amount to learning 2^{k+1} parameters.

The Naive Bayes classifier reduces the number of parameters to learn by assuming conditional independence, that is components of the random vector X are conditionally independent given C . Under this assumption, the posterior probability can be computed as,

$$P(X = x_q | C = c_i) = \prod_{j=1}^{k} P(X = x_{qj} | C = C_i) \tag{6}$$

Once the posterior probabilities are estimated, the most probable entity for a given query name X_q can be estimated as,

$$C \leftarrow \arg\max_i \, log(P(C = c_i)) + \sum_{j=1}^{k} log(P(X = x_{qj} | C = c_i)) \tag{7}$$

We used two different approaches for computing the posterior $P(X = x_{qj} | C = C_i)$.

Shingle Presence. Our first approach considers X as a binary random vector. Specifically, if X_n is an outcome of the random vector that represents a name n, then $X_{nj} \in \{0, 1\}$ depending on whether the shingle corresponding to the j^{th} component is not present or present in the name. Given this formulation, the posterior $P(X = x_{ni}|C = C_i)$ can be estimated by computing the frequencies of a shingle's presence for the given entity class,

$$P(X = x_{nj}|C = c_i) = \frac{\#\{x_j = 1 \cap C = c_i\} + \delta}{\#\{C = c_i\} + k\delta} \tag{8}$$

where δ is the Laplace smoothing factor and k is the total number of shingles derived from the training set.

Multinomial Model. For our second approach, instead of restricting each X_{nj} to be binary, we let X_{nj} to represent the number of times the shingle corresponding to the j^{th} component occurs in the given name. The posterior $P(X = x_n|C = C_i)$ can then be considered as the multinomial distribution, with the shingles as the different categories. The probability of occurence of a shingle given a entity class can then be estimated as,

$$P(X = x_{qj}|C = c_i) = \frac{\#\{x_j \cap C = c_i\} + \delta}{\sum_{j=1}^{k} \#\{x_i \cap C = c_i\} + k\delta} \tag{9}$$

where δ is the Laplace smoothing factor and k is the total number of shingles derived from the training set.

3.3 Experiments

Dataset. As mentioned, our approach was motivated by the VIAF collection and we used the same for empirically testing our approaches. The VIAF collection consists of name authority records for around 2.5 million entities. Each entity record in the VIAF collection consists of a set of names that could be associated to the entity and may also include other information. We eliminated most names in languages other than those in Western European Languages. We further filtered this dataset by removing entities with less than 4 names. These steps reduced our dataset to 291,952 entities and 1,886,049 names. Names were normalized by transforming to lower case and further by removing all punctuation except space. We retained space as shingles containing the space naturally modeled the components of the name. We also removed existence dates from the names as this information was available separately in the VIAF record of the entity corresponding to the name.

Experiment Design. To test our classifiers, we split our collection of name instances from the VIAF dataset into training and testing sets using a 70-30% split ratio. This gives 291,952 entity classes with 1,320,234 training instances and 565,815 testing instances. Each training and testing instance includes a name, the entity class it belongs to, and birth and death dates. We used a shingle length

of 3 characters to compute our feature representations, for our training set this amounted to 14,103 unique shingles.

Given a query name from the test set, we predict its entity class and compare it with the actual entity class the name belongs to. To compare our classifiers, we used the %accuracy measure. In addition to considering a prediction to be accurate if the correct entity is ranked first, we also considered the cases when the correct entity comes within top-5 or top-10 of the ranked list. We use a reduced search set by ranking entities based on how many shingles in the query name are indexed by them and considering only the top-M entities of the ranked list. For all our results below we used M = 20.

Results. Our first objective was to predict the entity a query test name belongs to using only the strings present in the name. Table 3 shows the accuracy results for our three approaches. The Edit distance based approach is attractive as no model estimation, other than building the shingle reverse index, is neccessary. However, it did not perform as well as the probabilistic approaches. The main reason was that edit distance is sensitive to the order in which the components of the name occur. For example, a sample of different names for the physicist Richard Feynman are shown in Table 1. A majority of the names start with

Table 1. Alternate names for the Physicist Richard Feynman

Alternate names for the Physicist Richard Feynman
Feynman, Richard P
Phillips Feynman, Richard
Feinman, Richard P
Feynman, Richard
Feynman, Richard Phillips
Feynman, R. P.
Feynman

the lastname Feynman followed by the firstname Richard. If the query name is 'Richard Feynman' the average edit distance between the query name and the names in the collection would be rather high than compared to a name such as 'Richard Gere'. The probabilistic approaches handle this issue as the shingle features capture information that is much more local than the entire string of the name. The multinomial model works better than the shingle presence model as it captures the frequency of a shingle's occurence, rather than its mere presence, in the names associated with an entity. Table 2 shows the top-5 results for the query name 'Richard Feynman' for the three approaches. The names shown were randomly sampled from the entity classes. Although the Multinomial Model is able to predict the correct entity with reasonable accuracy, for automatic name matching we would need the top-ranked entity to be correct with a much higher accuracy. This motivated us to use the existence dates as an additional information. To do this, we boosted the scores given to the entity classes by

Table 2. Top-5 results for the query name 'Richard Feynman'

Edit Distance	Shingle Presence	Multinomial Model
Norman, R. J	Calichman Richard F	Harman, Richard Michael...
Staar, Richard Felix	Harman, Richard Michael...	Calichman Richard F
Calichman Richard F	Babcock, Richard Felt	Feynman, Richard P
Manby, Richard,	Kern, Richard	Manby, Richard
Kern, Richard	Kahn, Richard	Kahn, Richard

Table 3. Accuracy results for automatic name disambiguation using only the name strings

Approach	**Edit Distance**			**Shingle Presence**			**Multinomial Model**		
Criteria	*First*	*top-5*	*top-10*	*First*	*top-5*	*top-10*	*First*	*top-5*	*top-10*
%Accuracy	42.9%	72.18%	83.4%	43.12%	74.87%	83.3%	60.82%	82.49%	86.71%

our classifiers if the query name's existence dates matched with the existence dates associated with the entity class. We used the following rules to boost the classifier scores.

$$score = \begin{cases} score + \gamma & \text{both dates match} \\ score + \gamma/2 & \text{either birth or death dates match} \\ score & \text{dates do not match} \end{cases}$$

Instead of checking if the dates exactly matched, we assumed two dates to match if the difference between the dates was less than 5 years. This, we felt, would handle cases when there were reasonable variations in the existence dates. Table 4 shows the top-5 results for the query name 'Richard Feynman' for the three approaches with the scores boosted using existence dates information. The names shown were randomly sampled from the entity classes. Table 5 shows

Table 4. Top-5 results for the query name 'Richard Feynman' using existence dates information

Edit Distance	Shingle Presence	Multinomial Model
Feynman, Richard P	Kahn, Richard F	Feynman, Richard P
Kahn, Richard F	Feynman, Richard P	Kahn, Richard F
Hamann-MacLean, Richard	Hamann-MacLean, Richard	Hamann-MacLean, Richard
Staar, Richard Felix	Babcock, Richard Felt	Babcock, Richard Felt
Babcock, Richard Felt	Staar, Richard Felix	Staar, Richard Felix

that the accuracy of our classifiers significantly improved. These results were obtained using $\gamma = 100$ Using this technique for merging the SNAC entities, with the entire VIAF dataset as the training collection, would simply require us to replace the Cheshire IR system with our classifiers towards finding common

Table 5. Accuracy results for automatic name disambiguation using name strings and existence dates

Approach	**Edit Distance**			**Shingle Presence**			**Multinomial Model**		
Criteria	*First*	*top-5*	*top-10*	*First*	*top-5*	*top-10*	*First*	*top-5*	*top-10*
%Accuracy	80.8%	89.14%	89.47%	84.72%	89.47%	89.49%	80.21%	89.25%	89.49%

VIAF records linking a pair of SNAC entities. However, we wanted to rigorously evaluate the match accuracies for the entire production SNAC dataset. We are currently investigating sampling techniques towards this purpose. We are also investigating extending our multinomial approach using ngram language models towards modeling name variants associated with an entity.

4 The Prototype SNAC Interface

The prototype public interface for the SNAC EAC-CPF database was developed by the California Digital Library and uses their open source eXtensible Text Framework (XTF) system to support search, display and navigation of the EAC-CPF records.

As shown in Figure 1 the user is able to browse the people, organizations, and families in the SNAC collection alphabetically by name, by occupation (for persons) or by subjects associated with the source EAD collection. Some interesting example records are also made available as "Featured Records" on this initial screen, which also provides a search capability that can access any of the content of the records (with names weighted more highly in the results).

Figure 2 shows a single record for the American General George S. Patton. The prefered form of the name for entries is shown (Alternate forms of the name from library authority files are available as a pop-up window). The record also shows (on the left-hand side) the occupations, and topical subject headings associated with the entity. If one or more of the source EAD records includes a biographical entry it is included also. On the right-hand side of the page access is provided to the archival collections created by the named entity (creatorOf), or in which the entity is referenced (referencedIn). The same panel provides access to other people and corporate bodies associated with the named entity. These are often the correspondents, family members or others referenced in the EAD sources. This data is also the basis for the constructing the social network of the named entity as shown in Figure 3 (this graphical depiction of the social network is accessed through the "graph demo" link at the top right). Resources such as books by or about the named entity are also included as are linked data resources, such as links to the library authority record for the named entity.

5 Conclusions

This paper has described the Social Networks and Archival Context project, and examined some of the issues and processes in deriving an authority database for

archival collections from EAD records using the EAC-CPF format. We examined some of the issues in matching and merging named entities from different collections and reported on some of our experimental work in name matching. Finally we described the prototype interface for public use of the SNAC database generated by the processing, matching and merging of named entities.

The SNAC project is still in progress, and we are now starting to apply the results of our research (and our ongoing analysis of matching and merging failures) to improved versions of the database.

Acknowledgments. The work presented in this paper is based on the Social Networks and Archival Context project (Daniel Pitti, Principal Investigator) funded by the (U.S.) National Endowment for the Humanities. Information on the SNAC project, and access to the prototype is available at http://socialarchive.iath.virginia.edu/ .

References

1. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 17th International Conference on Computational Linguistics, vol. 1, pp. 79–85 (1998)
2. Bunescu, R., Pasca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of EACL, p. 6 (2006)
3. Larson, R.R., McDonough, J., O'Leary, P., Kuntz, L., Moon, R.: Cheshire II: Designing a next-generation online catalog. Journal of the American Society for Information Science 47(7), 555–567 (1996)
4. Mann, G.S., Yarowsky, D.: Unsupervised personal name disambiguation. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, vol. 4, pp. 33–40 (2003)
5. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 127–136. Springer, Heidelberg (2001)

How to Become a Group Leader? or Modeling Author Types Based on Graph Mining

George Tsatsaronis[1], Iraklis Varlamis[2], Sunna Torge[1], Matthias Reimann[1], Kjetil Nørvåg[3], Michael Schroeder[1], and Matthias Zschunke[1]

[1] Biotechnology Center (BIOTEC), Technische Universität Dresden, Germany
george.tsatsaronis@biotec.tu-dresden.de
[2] Dept. of Informatics and Telematics, Harokopio University of Athens, Greece
[3] Dept. of Computer and Information Science, NTNU, Norway

Abstract. Bibliographic databases are a prosperous field for data mining research and social network analysis. The representation and visualization of bibliographic databases as graphs and the application of data mining techniques can help us uncover interesting knowledge regarding how the publication records of authors evolve over time. In this paper we propose a novel methodology to model bibliographical databases as *Power Graphs*, and mine them in an unsupervised manner, in order to learn basic author types and their properties through clustering. The methodology takes into account the evolution of the co-authorship information, the volume of published papers over time, as well as the impact factors of the venues hosting the respective publications. As a proof of concept of the applicability and scalability of our approach, we present experimental results in the *DBLP* data.

Keywords: Power Graph Analysis, Authors' Clustering, Graph Mining.

1 Introduction

Currently, vast amounts of scientific publications are stored in online databases, such as *DBLP* or *PubMed*. These databases store rich information such as the publications titles, author(s), year, and venue. Less often they provide the abstract, or the full publications' content and references. The exploitation of additional features, such as co-authorship information, may help us create novel services for bibliographic databases.

In this direction, new online services that process metadata have appeared, such as *ArnetMiner* [9] or *Microsoft Academic Search*[1]. Services that visualize co-authorship information are also available, such as the *"Instant graph search"*[2], which presents the existent co-authorship paths connecting two authors, or the *"Social graph"*[2], which presents all the co-authors of a single author in a star topology. However, to the best of our knowledge, there is currently no methodology available that models the evolution of the authors' publication profile and

[1] http://academic.research.microsoft.com/
[2] Co-author Path and Graph in Microsoft Academic Search.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 15–26, 2011.

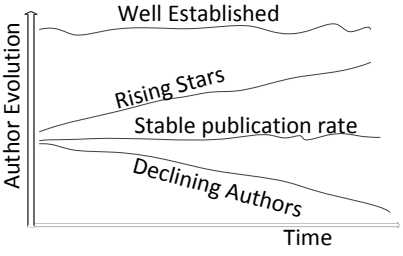


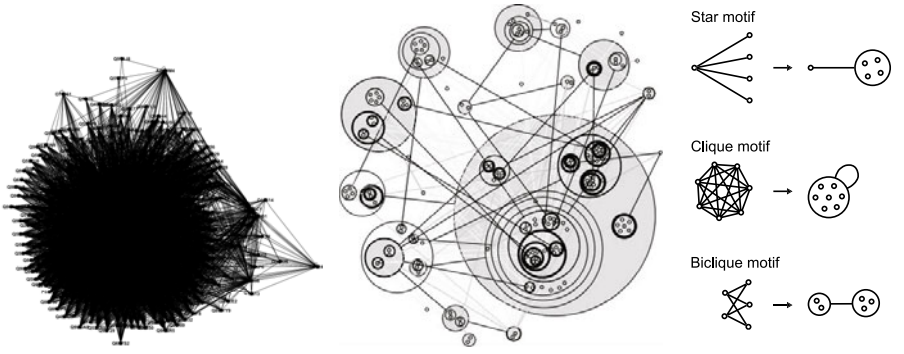
Fig. 1. Motivation: Four basic author "evolution motifs" over time

detects different types of authors in a bibliographical database, based on their evolution and standing over time.

Figure 1 presents, in a simplified manner, four intuitive author evolution motifs over time. We first distinguish between authors with an *ascending behavior* (*rising stars*), who show high increase in the amount and impact of their published work, as well as on the amount of collaborations with other researchers, and authors with a *descending behavior* (*declining authors*), who have a declining rate and impact of publications. Furthermore, in case the evolution is more *static*, we distinguish between authors that produce constantly a large amount of work over that time frame (*well established*) and authors that produce fewer publications, in a lower but stable rate (*stable publication rate*).

Motivated by the aforementioned motifs, this work addresses the author modeling problem using unsupervised learning, since supervised learning is not applicable due to the absence of manually annotated data for this task. First, we define four basic features that capture the authors' publication profile. Secondly, we monitor the evolution of these features over time and generate respective *evolution indices* per author. Finally, we use these indices to cluster authors with similar evolution profile into respective groups. The exact properties of each author evolution motif are deduced from the analysis of the final clusters.

Our methodology introduces two novel ideas. The first is the application of *Power Graph Analysis* [5] to the co-authorship graphs constructed from bibliographical data, which is performed for the first time, to the best of our knowledge, in bibliographical analysis. The use of *Power Graphs* allows fast and large-scale clustering experiments since they can compress by even up to 40% the information of the original co-authorship graph, as we show in our experiments, in a lossless information manner. Also, they can visualize more efficiently than the original graphs the co-authorship information by identifying several motifs, e.g., cliques and bi-cliques. The second is the introduction of a set of features for each author, which is based on the *Power Graphs* structure, the number, and the impact of her publications. Through the features' monitoring over time, the respective *evolution indices* are computed, based on which the authors' clustering is conducted. Finally, the indices are employed in the analysis of the resulting clusters as descriptors of the *authors' dynamics*. The contributions of this work can, thus, be summarized into the following: (a) a novel methodology for



(a) Huge biological "fur ball" network. (b) After the application of *Power Graphs*. (c) Basic motifs recognized by *Power Graph Analysis*.

Fig. 2. Figure 2(a) shows an example of a huge biological network. Figure 2(b) shows the corresponding *Power Graph*. The three basic motifs recognized by *Power Graphs* are shown in Figure 2(c): *Star*, *Clique* and *Biclique*. *Power Nodes* are sets of nodes and *Power Edges* connect *Power Nodes*. A *Power Edge* between two *Power Nodes* signifies that all nodes of the first set are connected to all nodes of the second set.

modeling the *dynamics* of authors' publication profiles, and clustering them into groups, (b) transfer of the *Power Graph Analysis* methodology from the field of bioinformatics, to the field of bibliographical databases analysis, in order to visualize and process co-authorship graphs, and, and, (c) empirical analysis and demonstration of the applicability of our approach in a large bibliographical data set (*DBLP*). The rest of the paper is organized as follows: Section 2 presents some preliminary concepts and discusses related work. Section 3 introduces our methodology and in Section 4 we present our experimental findings. Section 5 concludes and provides pointers to future work.

2 Preliminaries and Related Work

2.1 Visualizing Graphs with Power Graphs

In the bioinformatics field, networks play a crucial role, but their efficient visualization is difficult. Biological networks usually result in *"fur balls"*, from which little insight can be gathered. In the direction of providing an efficient methodology for visualizing large and complex networks, such as protein interaction networks, the authors in [5] introduce *Power Graph Analysis*, a methodology for analyzing and representing efficiently complex networks, without losing information from the original networks. The analysis is based on identifying *re-occurring network motifs* using several abstractions. The three basic motifs recognized by *Power Graphs* are shown in Figure 2. These are the *Star*, the *Clique* and the *Biclique*, and constitute the basic abstractions when transforming the original

graph into a *Power Graph* with *Power Nodes*, i.e., sets of nodes, connected by *Power Edges*. *Power Graphs* offer up to 90% compression of the original network structure [5], allowing for efficient visualization. Figure 2 shows an example of a *"fur ball"* network, and its transformation after the application of *Power Graph Analysis*.

Power Graphs have been successfully applied in bioinformatics, as the networks are rich in the aforementioned motifs [5]. Co-author networks in the bibliographic analysis are implicitly built on such motifs and, thus, perfectly suited for applying *Power Graph Analysis*. A publication is either considered as *Clique* of all authors or as *Biclique* with first and last authors on one and all other authors on the other end. Motivated by this, in this paper we apply *Power Graphs Analysis* into co-authorship graphs extracted from bibliographical databases, in order to define authors' features. As shown in the experimental section, the resulting *Power Graphs* allow for a very efficient visualization of the co-authorship graph.

2.2 Mining Graphs from Bibliographical Databases

Graph-based mining methods in bibliographic databases usually create a graph from author names, venues, or papers' topics, apply a graph partitioning algorithm to locate interesting sub-graphs, and present results in the form of node clusters, e.g., authors by topic, or through visualization of the graph, e.g., co-authors of a single author in a star topology. The various methods for representing bibliographical databases as graphs, can be divided in two categories: (i) those that use n-partite graphs, which contain for example authors, conferences, or topics as nodes, and edges that connect different node types representing relations, and, (ii) those that use graphs with a single node type and edges that may vary in meaning depending on the application. An example of the former category can be found in [8], where bipartite models connecting conferences to authors are employed to rank authors and conferences. Tripartite graph models for authors-conferences-topics have also been introduced in the past [9]. In the later category, e.g., the work in [2], nodes correspond to authors, and edges represent citation or co-authorship relation. The resulting representation can be used to rank authors, find author communities, measure *author centrality* [3,4,6], or find special relations between authors, such as advisor-advisee [10]. Regarding the evolution analysis of graphs, snapshot-based approaches, e.g., an author-paper graph per year, are frequently used [1,8]. In these approaches, predefined measures from each snapshot are extracted and monitored over time. In this work we present an approach which differs from the aforementioned in several points; (a) the co-authorship graph is used only as a basis for defining collaboration-related authors' features , (b) the authors' clustering process analyzes the evolution of these features over time, as well as the changes in the volume and impact of the authors' publications, and, (c) the clustering aims at identifying author evolution motifs and uncover their characteristics, and not to detect author communities.

Paper ID	Authors
p1	a1, a3
p2	a1, a4
p3	a1, a5
p4	a2, a3
p5	a2, a4
p6	a2, a5
p7	a3, a6
p8	a3, a7
p9	a3, a8
p10	a3, a9
p11	a3, a10
p12	a20, a31
p13	a20, a32
p14	a21, a31
p15	a21, a33
p16	a22, a31
p17	a22, a34

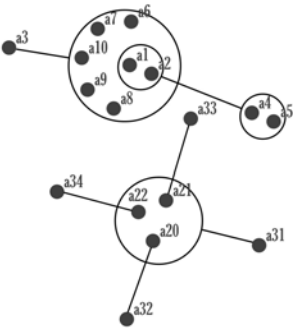


Fig. 3. A sample co-authorship *Power Graph*

3 Approach

The suggested methodology comprises the following steps: (1) creation of the co-authorship *Power Graphs* in different time points, i.e., years, within a given time frame (Section 3.1), (2) computation of each feature per time point, and of each features' *evolution index* per author (Section 3.2), and, (3) clustering of authors based on their *evolution indices* (Section 3.3).

3.1 Co-authorship Graphs with *Power Graphs*

The initial co-authorship graph contains authors as nodes and weighted edges that connect pairs of authors. Given a paper with k authors, an edge connecting each pairwise combination of the k authors is added in the graph. Given a specific time point t_i (e.g. a year), the initial co-authorship graph accumulatively contains all the publication records from the beginning until t_i. Edges' weights represent the number of papers that the two authors have co-authored until t_i.

The original graph is converted to a *Power Graph* as explained in [5]. An example of bibliographic records (papers and authors), and the resulting co-authorship *Power Graph* is depicted in Figure 3, where for reasons of simplicity we assume that all papers have exactly two authors. Authors *a1* and *a2* have exactly the same co-authors (*a3*, *a4*, *a5*). The same holds for authors *a3*, *a4* and *a5*. This is depicted by a *bi-clique* in the *Power Graph*. Author *a3* has collaborated with *a1*, *a2* and *a6* to *a10*. So *a3* forms a *star* with his co-authors, who form a pair of nested *Power Nodes* (set *a1*, *a2* is inside the greater *Power Node*). Finally, all the co-authors of *a31* form a *star*. The transformation of the original co-authorship graphs into *Power Graphs* offers three very important advantages: (i) it performs a first-level clustering of the authors based on their co-authorship information, (ii) it compresses the original graph, without losing information, and, (iii) it allows for a more efficient visualization of the co-authorship information.

3.2 Authors' Features and *Evolution Indices*

After constructing the co-authorship *Power Graphs*, the next step is to define the features based on which authors' evolution may be measured. Given a *Power Graph* G_i in time point t_i, we define the following features for every author a_k: (1) the size of the *Power Node* to which a_k belongs (S_i) (if a_k is not member of a *Power Node* then S_i=0), (2) the sum of the *Power Nodes'* sizes with which a_k's *Power Node*, or any *Power Node* containing her *Power Node*, is connected (C_i), (3) the number of papers authored by a_k (P_i) until t_i, and, (4) the aggregated impact of a_k's publications until t_i (I_i)[3]. The intuition behind each of the aforementioned features is straightforward; S_i measures the number of the most frequent co-authors an author has at time point t_i (size of her *clique*), C_i measures the size of her extended *clique*, i.e., the co-authors of her co-authors, P_i measures the number of publications up to t_i, and, finally, I_i measures the total impact of her work.

More formally, for an author a_k who belongs to *Power Node* V_k, in the *Power Graph* for time point t_i, S_i is defined as:

$$S_{ik} = we_{V_k,V_k} \cdot |V_k| \tag{1}$$

where $|V_k|$ is the size of *Power Node* V_k, and we_{V_k,V_k} is the weight of the edge connecting *Power Node* V_k with itself. This later weight shows the *strength* of the clique formed by the authors in *Power Node* V_k. Let V_k be connected to n other *Power Nodes*. Then, C_i, for author a_k is defined as:

$$C_{ik} = \sum_{m=1..n} we_{V_k,V_m} \cdot |V_m| \tag{2}$$

where V_m is any *Power Node* connected to V_j with an edge of weight we_{V_j,V_m}. P_i is defined as the number of papers produced by author a_k up to time point t_i. If time points are years, then P_{ik} denotes the number of papers that author a_k has produced from the beginning of her career up to t_i. Finally, I_i is the sum of the impact factors of the authors publications up to t_i. More formally, if the author has written n papers up to t_i, then I_i is defined as follows:

$$I_{ik} = \sum_{m=1..n} IF_m \tag{3}$$

where IF_m is the impact factor of the venue or the journal where paper m was published.

The next step is to define an *Evolution Index* for each of these four features (S_{ik}, C_{ik}, P_{ik}, and I_{ik}), which capture the way the feature values evolve over time. For this reason, given a time span $T : [t1, t2]$, for which we monitor authors, we

[3] We assign to each paper the impact factor of the venue or journal, in which each paper was published. For our experiments we used the list maintained by *Citeseer*. Historical impact factors are not taken into account due to the lack of respective data.

build a *Power Graph* G_i for each $t_i \in T$ (e.g., for each year). Our aim is to capture the *dynamics* of author a_k in each of the four feature dimensions. For the definition of each of the four *Evolution Indices* we employ a function, which we call *change*. *Change* measures the ratio of the change of any of the features S_{ik}, C_{ik}, P_{ik}, and I_{ik} from time point t_{i-1} until time point t_i. *Change* for feature S_{ik} of author a_k is defined as:

$$Schange_{ik} = \frac{S_{ik} - S_{(i-1)k}}{S_{ik}} \tag{4}$$

The above equation captures the ratio of *change* occurred in a_k's *Power Node* from t_{i-1} to t_i. If a_k's clique has grown from t_{i-1} to t_i then the respective *Power Node* V_k size will increase, and $Schange_{ik}$ will be positive (i.e., in contrast to 0 where there is no change). In a similar manner, $Cchange_{ik}$ captures the ratio of the change in the author's connectivity with other cliques, $Pchange_{ik}$ captures the change with regards to the volume of the papers produced from t_{i-1} until t_i, and $Ichange_{ik}$ the change in her publications' impact.

We can now define the *Evolution Index* (*EI*) for the S feature (*S.EI*), given T, as shown in the following equation:

$$S.EI_{Tk} = \max_{t_i \in T} Schange_{ik} \cdot S_{t_2 k} \cdot \sum_{t_i \in T} Schange_{ik} \tag{5}$$

where t_2 is the final time point in the examined time frame T. Equation 5 captures the evolution of the author over the time frame T, and measures the *dynamics* of the author in the dimension of feature S, as it takes into account the maximum occurred change, the standing of the authors according to S in the final time point, and the sum of all occurred changes. Similarly, the $C.EI_{Tk}$, $P.EI_{Tk}$, and $I.EI_{Tk}$ *Evolution Indices* can be defined for C, P and I features respectively.

3.3 Clustering Authors Using *Bisecting K-Means*

In this section we demonstrate the use of the *Evolution Indices* in the author clustering task. The clustering algorithm that we employ is *bi-secting K-Means*, which delivers high clustering performance, better than *K-Means*, and other agglomerative techniques [7]. *Bisecting K-Means* starts with a single cluster containing all data points, and iteratively selects a cluster and splits it into two clusters until the desired number of clusters is reached or a quality criterion *coherence* is met. Its time complexity is linear to the number of data points.

In our case, each author a_k is a single data point with four dimensions, which correspond to the four evolution indices. Finding authors' clusters is, consequently, formulated as a typical clustering problem, which can be solved using *bisecting K-Means*. Authors are represented as vectors in the four dimensional

Input: Database of papers D, time frame $T[t_1, t_2]$, number of desired clusters k
Output: A clustering solution of authors into k groups

Construct *Power Graph* at t_0, i.e., the previous time point from t_1
foreach *time point* $t_i \in T$ **do**
    Construct *Power Graph* at t_i
    foreach *author* $a_k \in D$ **do**
        Measure and store $Schange_{ik}, Cchange_{ik}, Pchange_{ik}, Ichange_{ik}$
Measure and store $S.EI_{Tk}, C.EI_{Tk}, P.EI_{Tk}, I.EI_{Tk}$
Put all authors $a_k \in D$ into a single cluster
Pick a cluster to split
Find 2 sub-clusters using the basic *K-Means* algorithm (bisecting step)
Repeat step 9 for *ITER* times and choose the best split
Repeat steps 8, 9, and 10 until number of clusters is k

Algorithm 1. Clustering authors of a bibliographical database using *Power Graphs* and *bisecting K-Means*.

space, so the *cosine similarity* measure can be used for measuring similarity between two data points, or between a data point and a cluster centroid. The centroid of a cluster K of authors ($\boldsymbol{C}$), is defined as follows:

$$\boldsymbol{C} = \frac{1}{|K|} \sum_{a_k \in K} \boldsymbol{a_k} \tag{6}$$

Algorithm 1 describes our methodology of organizing authors in a bibliographic database into k groups. If the bibliographical database contains m papers written by n distinct authors, and the resulting *Power Graph* at any time point contains maximum pn *Power Nodes*, then the complexity of the algorithm can be summarized into $O(|T| \cdot (m + n^2 log n + pn) + n)$, which, even for millions of data points, makes the task computationally feasible.

4 Evaluation and Results

In order to demonstrate the organization of authors into categories, and analyze the properties, we experiment using our methodology with the *DBLP Computer Science Bibliography* database. The database comprises 925,324 distinct author names, and 1,601,965 publication entries (papers). For our experiments we select the authors, and their publications, that have in total a minimum of 5 publications by 2010. We then apply our methodology using two different experimental set-ups: (a) we use time frame $T = [2000, 2010]$ to cluster the authors into four categories, and analyze the clusters' properties, and, (b) we use time frame $T = [2000, 2005]$ and examine whether our method organizes successfully authors into clusters, by evaluating the behavior of each cluster in the next five years, i.e., [2006, 2010]. For our experiments we construct the *Power Graphs* from the original co-authorship graphs, for the periods [2000, 2010], and [2000, 2005] respectively. Table 1 reports statistics from the construction of the *Power Graphs* for all years in [2000, 2010], where the edge reduction rates reached up to 41%.

Table 1. Number of nodes and edges in the original co-authorship graphs per year, and in the constructed *Power Graphs*

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
#Nodes	53,679	60,732	67,530	75,490	83,797	92,453	100,301	108,051	113,548	117,619	119,767
#Edges	146,562	174,951	207,691	244,132	285,517	331,618	377,586	427,991	473,601	518,158	552,956
#Power Nodes	22,396	25,634	28,730	32,584	36,640	40,972	45,040	49,069	52,060	54,414	56,046
#Power Edges	87,346	103,864	122,279	143,698	168,736	197,421	227,889	261,704	293,246	323,583	349,538
#Edge Reduction Rate	0.404	0.406	0.411	0.411	0.409	0.404	0.396	0.388	0.38	0.375	0.367

Cluster Centroids

Cluster ID	Feature S	Feature C	Feature P	Feature I
1	29.96	7.16	35.25	14.32
2	64.01	12.8	18.17	8.18
3	6.12	1.37	31.82	6.29
4	5.56	11.21	23.63	13.79

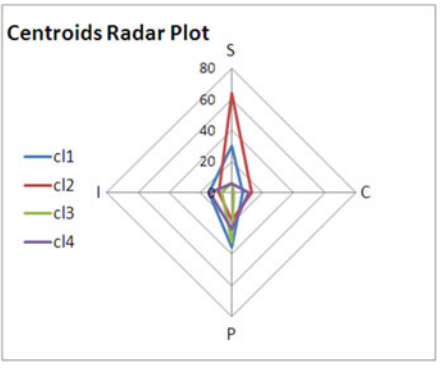
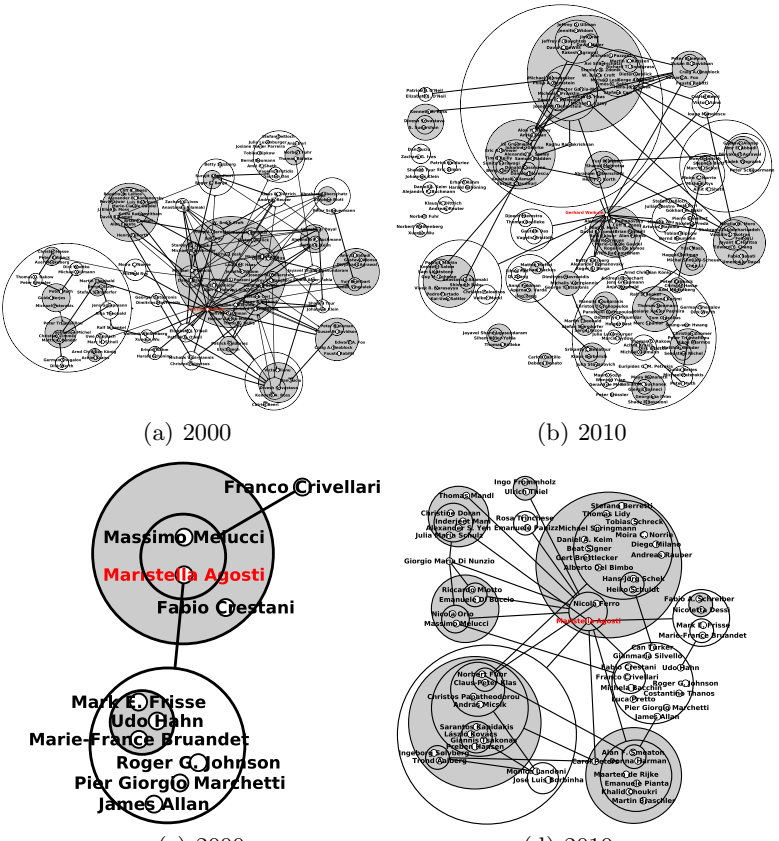


Fig. 4. The centroids of the four basic clusters, and their radar plot according to the four feature dimensions

4.1 Clustering *DBLP* Authors

In Figure 4 we show the results of experiment (a). The number of author clusters was set to 4, in an effort to identify the basic intuitive motifs shown in Figure 1. We performed a series of experiments with an increasing number of *ITER* in *bi-secting K-means* (from 5 to 500). As expected, higher values produced more stable clusters. The left part of the figure shows the centroids of the four clusters (cluster 1 to 4), and their respective values in the four dimensions (S, C, P, and I). The right part of the figure shows a radar plot of the four centroids in the four different dimensions. Each polygon represents a cluster centroid, and expands more towards a specific direction, if the respective feature value is high.

Cluster 2 contains the most *connected* authors (highest S). The majority of the authors in this cluster are *well established* with great dynamics in expanding their collaborations (C). The explanation is that most of the authors in cluster 2 are group leaders or professors and have many collaborations. One can certainly trace *rising stars* inside that cluster, since its points have huge dynamics in collaborations. Some examples of authors in that cluster are *Christos H. Papadimitriou*, and *Maristella Agosti*. Cluster 1 is certainly the group with the most candidate *rising stars*. The authors in this cluster show the best dynamics in paper publishing (P), and also really good dynamics publications' impact (I).



(a) 2000 (b) 2010

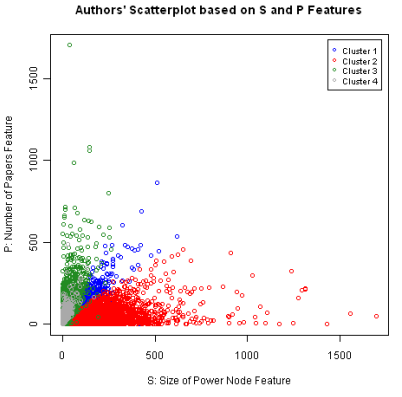
(c) 2000 (d) 2010

Fig. 5. Evolution of *Power Graphs* for two authors over two time points. Grayed circles denote a *Clique* and white a *Biclique*.

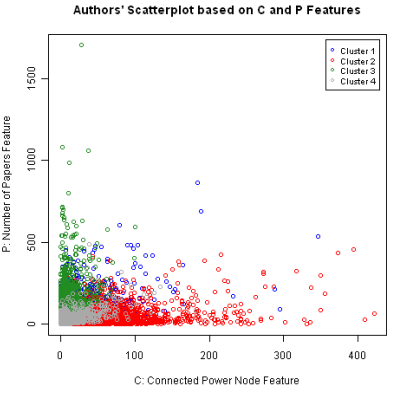
Some of the authors in cluster 1 are already well established, but the common characteristic of all authors in cluster 1 is their high dynamics in three dimensions (S, P, and I). Some examples of authors in cluster 1 are *Gerhard Weikum*, and *George Buchanan*. The third most interesting cluster is 4. Authors here certainly publish much and have great potential, but they still need to work their I and S features. Clusters 3 and 4 contain stable publishing authors, without special dynamics though. Finally, in cluster 3 the main motif is that of isolated authors (low S and C values). This cluster also contains *declining authors*, which have ceased expanding their collaborations, but their dynamics in paper publishing (P) remain high.

In Figure 5 we show an example of the *Power Node* evolution between 2000 and 2010 for two authors: *Gerhard Weikum* from cluster 1, and *Maristella Agosti* from cluster 2[4]. The figure also shows their connected *Power Nodes*. Figures 5a

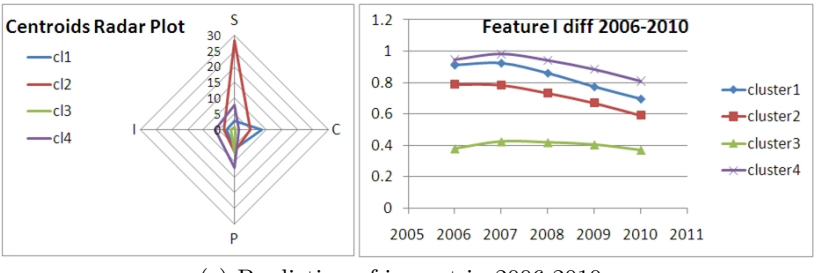
[4] Zoom is possible in the electronic edition.



(a) S and P Features.



(b) C and P Features.



(c) Prediction of impact in 2006-2010.

Fig. 6. Figures 6(a) and 6(b): Authors' scatter plots based on two feature combinations. Figure 4.1: Clustering authors in 2000-2005, and predicting impact in 2006-2010.

and b, show how S and C evolved for *Gerhard Weikum*, and Figures 5c and d, for *Maristella Agosti*: the author of cluster 2 had larger changes in her *Power Node* and its neighborhood, compared to the author of cluster 1. This example demonstrates the general motif of clusters 1 and 2 in the radar plot of Figure 4 regarding S and C.

In Figures 6a and 6b we show the authors' scatter plot based on two feature combinations; S and P, and P and C. The figure demonstrates an example of the features' ability to separate the authors. As shown, S and P can separate cluster 1 from the rest, while C and P can separate cluster 3 from 4. Similar findings were observed in all the remaining features' pairs. Figure 6c shows the results of experiment (b). The left part shows the radar plot of the authors' cluster centroids for $2000-2005$. The clustering predicts that authors in cluster 4 have large dynamics in increasing their publications' impact factor (I). The right part shows the yearly increase in authors' impact factor for $2006-2010$. It verifies that authors in cluster 4 had the largest difference (on average over all cluster points) per year on the impact feature compared to the authors of the rest clusters.

5 Conclusions

In this paper we introduced a novel methodology for the organization of authors into basic clusters, using *Power Graph Analysis*. We defined evolution indices over features that capture the connectivity and strength of the authors' co-operations, as well as their publications' volume and impact over time. We demonstrated the applicability of our approach to capture the dynamics of authors using the evolution of the four defined features by clustering authors in *DBLP* with *bi-secting K-Means*. It is in our next plans to explore and interpret authors' clustering using several different values for the k parameter. We also plan to explore the application of our methodology for comparing institutions, based on the notion of the centroid of the institutions' authors, as well as comparing scientific venues, or individual authors. In this direction, the comparison methodology would consider the publications of the respective institutions, venues, or authors, and follow the methodology in this paper. Similar entities, e.g., institutions, would result in similar clusters, for the same value of k, and the comparison could be feasible by placing the clusters at the same radar plot.

References

1. Erten, C., Harding, P.J., Kobourov, S.G., Wampler, K., Yee, G.: Exploring the computing literature using temporal graph visualization. In: Visualization and Data Analysis, pp. 45–56 (2003)
2. Ke, W., Borner, K., Viswanath, L.: Major information visualization authors, papers and topics in the acm library. In: INFOVIS, pp. 216.1–216.9 (2004)
3. Li, X., Foo, C., Tew, K., Ng, S.: Searching for rising stars in bibliography networks. In: Zhou, X., Yokota, H., Deng, K., Liu, Q. (eds.) DASFAA 2009. LNCS, vol. 5463, pp. 288–292. Springer, Heidelberg (2009)
4. Nascimento, M.A., Sander, J., Pound, J.: Analysis of sigmod's co-authorship graph. SIGMOD Rec. 32, 8–10 (2003)
5. Royer, L., Reimann, M., Andreopoulos, B., Schroeder, M.: Unraveling protein networks with power graph analysis. PLoS Computational Biology 4(7) (2008)
6. Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K., Sødring, T.: Analysis of papers from twenty-five years of sigir conferences: what have we been doing for the last quarter of a century? SIGIR Forum 36, 39–43 (2002)
7. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining, pp. 109–110 (2000)
8. Sun, Y., Wu, T., Yin, Z., Cheng, H., Han, J., Yin, X., Zhao, P.: Bibnetminer: mining bibliographic information networks. In: SIGMOD 2008, pp. 1341–1344. ACM, New York (2008)
9. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: KDD, pp. 990–998 (2008)
10. Wang, C., Han, J., Jia, Y., Tang, J., Zhang, D., Yu, Y., Guo, J.: Mining advisor-advisee relationships from research publication networks. In: KDD, pp. 203–212 (2010)

Find, New, Copy, Web, Page - Tagging for the (Re-)Discovery of Web Pages

Martin Klein and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA 23529
{mklein,mln}@cs.odu.edu

Abstract. The World Wide Web has a very dynamic character with resources constantly disappearing and (re-)surfacing. A ubiquitous result is the "404 Page not Found" error as the request for missing web pages. We investigate tags obtained from Delicious for the purpose of rediscovering such missing web pages with the help of search engines. We determine the best performing tag based query length, quantify the relevance of the results and compare tags to retrieval methods based on a page's content. We find that tags are only useful in addition to content based methods. We further introduce the notion of "ghost tags", terms used as tags that do not occur in the current but did occur in a previous version of the web page. One third of these ghost tags are ranked high in Delicious and also occurred frequently in the document which indicates their importance to both the user and the content of the document.

1 Introduction

The World Wide Web is a highly dynamic information space where resources very frequently surface, disappear and move from one location to another. As one result users often encounter a "404 Page Not Found" response when requesting a web resource. This error may occur when re-visiting a bookmark created some time ago, requesting a no longer valid URI or following a link from a badly maintained web page. Even though we know how to create URIs that do not change [3] there are many reasons why URIs or even entire websites break [13].

In previous work [8,9] we have provided two content based approaches to generate search engine queries that rediscover such missing pages. We are following our intuition that information on the web is rarely completely lost, it is just missing and we can utilize the web infrastructure (search engine, their caches, archives, etc) to rediscover and preserve these resources. Our previously introduced methods are the web page's title and the page's lexical signature. Both have shown to perform very well for the purpose of rediscovering missing web pages. However, both methods are applicable only if an old copy of the missing

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 27–39, 2011.

page can be found in the web infrastructure. If that fails we have no means to gain knowledge of the "aboutness" of the missing page.

As a third option we are motivated to investigate the retrieval performance of tags left by Delicious users to annotate URIs. We see several intriguing aspects for using tags: Unlike titles and lexical signatures tags may be available even if no old copy of a missing page can be found. That means even if we can not obtain the title or generate the lexical signature of the missing page we may find tags describing its content. Tags are created by many users, therefore somewhat utilize the "wisdom of the crowd". They have been predicted to be useful for search [4,5] and shown to possibly contain terms that do not occur in the original (now missing) web page. This can be beneficial for retrieving other, potentially relevant documents. We do not expect tags to outperform titles and lexical signatures but we foresee an added value for the rediscovery of missing web pages in combination with the previously established methods. In previously generated corpora containing randomly sampled URIs we experienced that tags were very sparse. In [9] for example we only found tags for 15% of all URIs. This led us to the creation of a new, "tag-centric" corpus introduced here. In summary, this paper's contributions are:

- determining the best performing tag based query length in number of terms
- analyzing the similarity and relevance of tag based search results
- quantifying the increased retrieval performance for a combination of query methods
- identifying tags as ghosts of pages that have past.

2 Related Work

2.1 Tags for Search

A lot of work has been done to investigate the usefulness of tags for search. Morrison [6] for example found in an extensive study that search in folksonomies can be as precise as search in major modern web search engines. By comparing Delicious data with search engine log data Krause et al. [12] found that tags and search terms both follow a power law distribution. That implies a low overall overlap and an increased overlap for the more frequent terms. They further found sparse overlap in Delicious and search engine rankings but if there was overlap it occurred at the top end of the rankings. Heymann et al. [5] conducted the probably most extensive study on tags with a dataset of about 40 million bookmarks from Delicious. Their results show that about half of the tags occur in the content of the page they annotate and 16% even occur in the page's title. Interestingly they found that in one out of five cases the tags neither occur in the page nor in the page's in- or outlinks. They conjecture that tags therefore can provide data for search that is otherwise not available. However, they state that annotated URIs are rather sparse compared to the size of a modern search

engine's index. Bischoff et al. [4] confirm the findings of Heymann et al. with almost 45% of their tags found in the page's text. Their user study shows that tags are mostly reliable and accurate and they are partially used the same way as search terms. Yanbe et al. [16] propose a social bookmarking-based ranking and use it to enhance existing link-based ranking methods. They also find that tag proportions stabilize over time which means users eventually come to an agreement over tags and even copy each other. The work done by Bao et al. [2] incorporates the frequency of tags used to annotate URIs as an indicator for its popularity and quality.

2.2 Content and Link Based Methods to Rediscover Web Pages

Content based search engine queries can be a powerful tool to rediscover missing web pages. We have shown in previous work [8] that lexical signatures are suitable. We found that 5− and 7−term lexical signatures perform best depending on whether the focus is on obtaining the highest mean rank (5 terms) or the most top ranked results (7 terms). Sugiyama et al. [14] have shown that the content of in- and outlinks of a web page can help refine the lexical signature of that page. We have built on that idea in [11] and determined optimal parameters to create a link neighborhood based lexical signature. Even though they are expensive to compute, similar to tags, they may provide an alternative if no copy of a missing page can be found in the web infrastructure. Further research in [9,10] has shown that titles of web pages are a very strong alternative to lexical signatures. The results also prove that we can increase the retrieval performance by applying both methods combined.

3 Experiment Setup

3.1 Data Gathering

We have seen in previous work [9] that for datasets based on randomly sampled URIs tags are very sparse and it is hard to aggregate a somewhat representative corpus. Heymann et al. [5] supports this point by showing that compared to a search engine's index the number of URIs annotated with tags is diminishing. Therefore we decided to reverse the approach and obtain tags and the URIs they annotate instead of first sampling URIs and then asking for their tags hoping to get a good sized sample set. Note that these URIs are not really missing but due to the sparseness of tags we use the obtained URIs and pretend they are missing. A few sources are available to obtain tags left by users to annotate URIs. The website `delicious.com` is probably the most famous and most frequently used one. We queried Delicious for 5000 unique URIs from their index using the Delicious "random tool"[1]. We are aware of the bias of our dataset towards the Yahoo! index (which we query against) especially in the light of Yahoo!

[1] `http://www.delicious.com/recent/?random=1`

Table 1. Tag Distribution

# of Tags	0	1-5	6-10	11-15	16-20	21-25	26-29	30
Frequency	0.42	1.44	2.36	4.48	6.66	6.86	4.04	**73.73**

integrating Delicious data into their index[2]. However, sampling from Delicious is an approach taken by various researchers [4,5].

We eventually aggregated 4968 unique URIs from Delicious. We did get 11 duplicates and despite the fact that we sampled from the "random tool" which pulls from the Delicious index we obtained 21 URIs that did not have tags. We used screen scraping, instead of the Delicious API, to gather up to 30 tags per URI[3]. The order, which may be of relevance for web search, indicates the frequency of use for all tags. Table 1 shows the relative distribution by number of tags for all URIs. We obtain the maximum of 30 tags for almost three out of four URIs.

3.2 Performance Measure

We use the Yahoo! BOSS API for all queries and analyze the top 100 results. We apply three different performance measures for our evaluation. Since our data corpus consists of live URIs one way of judging the performance of tag based search queries is to analyze the result set and monitor the returned rank of the URI of interest. This establishes a binary relevance case. More precisely, similar to our evaluation in [9] the first performance measure distinguishes between four retrieval cases where the returned URI is:

1. top ranked
2. ranked 2-10
3. ranked 11-100
4. considered undiscovered (ranked 101+).

We consider URIs not returned within the top 100 as undiscovered. We are aware of the possibility of discriminating against results returned just above that threshold but it is known that the average user does not look past the first few search results ([1,7]) which encourages our threshold. We also compute normalized Discounted Cumulative Gain (nDCG) for the result set as a measure to reward results at the top of the result set and penalize results at the lower end. We give a relevance score of 1 for an exact match of the target URI and a score of 0 otherwise. For comparison reasons we also include mean average precision (MAP) scores for our results with the same binary relevance scoring.

[2] http://techcrunch.com/2008/01/19/delicious-integrated-into-yahoo-search-results/

[3] We have previously shown the Delicious API to be unreliable, see: http://ws-dl.blogspot.com/2011/03/2011-03-09-adventures-with-delicious.html

We secondly compute the Jaro-Winkler distance between the original URI and the top ten returned URIs from the result set. The intuition is that some highly relevant pages have very similar URIs. The Jaro-Winkler distance is frequently applied to measure the similarity between short string such as names. It is therefore well fitting for comparing our URIs.

As a third measure we compute the Dice coefficient between the content of the original page and the content of the top ten search results. This gives us a sense of the string based similarity between the original content and the returned results. A high coefficient means a high similarity which in turn can be interpreted as a high relevance to the query - the tags used to annotate the original URI.

4 Retrieval Performance of Tags

4.1 Length of Tag Based Search Queries

We determined the best performing lexical signature length in previous work [8] to be 5 and 7 terms and initially assumed these parameters could be equally applied to tags. Hence we created queries consisting of 5 and 7 tags and issued them against the API. It turns out our assumption was inaccurate and therefore we widened the spectrum. Table 2 shows query lengths varying from 4 to 10 tags and their performance in relative numbers with respect to our four retrieval categories introduced in Section 3.2 as well as their nDCG and MAP. The generally low mean nDCG and MAP values are due to the large number of undiscovered URIs. Table 2 shows that 8−tag queries return the most top ranked results (11%) and 7−tag queries, tied with 6−tag queries, leave the fewest URIs undiscovered. Is also shows that 7− and 8−tag queries are tied for the best mean nDCG while 8 tags have a slight edge at MAP. However, taking this data we can not find a statistical significance (p-value ≤ 0.05) between the performances of 5−, 6−, 7− and 8−tag queries. The performance of 4−, 9− and 10−tag queries is in comparison statistically significantly worse.

Table 2. Relative Retrieval Numbers for Tag Based Query Lengths, nDCG and MAP

# of Tags	Top	Top10	Top100	Undis	Mean nDCG	MAP
4	7.2	11.3	9.6	71.9	0.14	0.11
5	9.0	11.3	**9.7**	69.7	0.16	0.13
6	9.7	**12.0**	9.0	**69.3**	0.17	0.14
7	10.5	11.5	8.7	**69.3**	**0.18**	0.14
8	**11.0**	10.8	8.1	70.1	**0.18**	**0.15**
9	10.3	9.9	8.0	71.9	0.17	0.14
10	9.7	8.9	6.4	75.0	0.15	0.13

4.2 Relevance of Results

Our binary retrieval evaluation (the URI is either returned or not) is applicable since we know what the "right" result to the tag based query is - the URI. However, the results in Table 2 indicate that a large percentage of URIs remain undiscovered. We are now investigating the relevance and similarity of the returned results for cases where the URI of interest is not returned.

We compute the Jaro-Winkler distance between the original URI and the URIs of top ten results to determine the similarity between URIs. Given the data from Table 2 we take the results of the five best performing tag based query lengths (5, 6, 7 and 8 tags) for this analysis. Figure 1 shows in the left graph the mean Jaro-Winkler distance for all URIs (y-axis) per rank (x-axis). Even though the four lines differ by point character (with respect to their length) it seems insubstantial to distinguish between them. The mean Jaro-Winkler value is high. It varies between 0.59 and 0.62 with slightly higher values for the top two ranks. The values for ranks three through ten are almost indistinguishable. These results show very similar URIs in the top ten indicating a high degree of relevancy for the returned results.

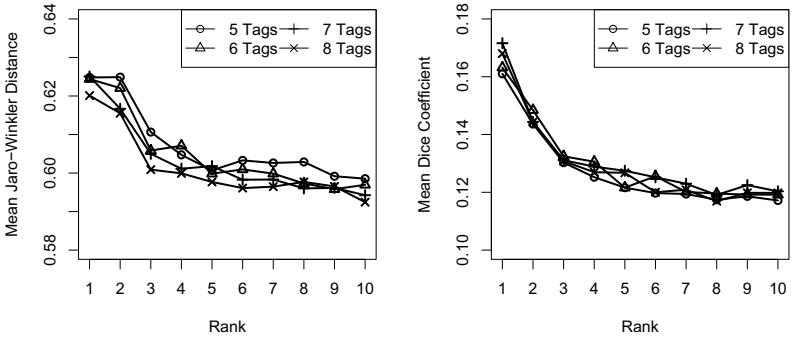


Fig. 1. Similarity Between URIs and Contents

Figure 1 shows in the right graph the Dice coefficient between the content of the URI the tags were derived from and the content of the top ten results. The intuition is that tags may not have the specificity to reliably return their URIs but contain enough information to return other relevant pages. This can especially be true for tags that do not actually occur in the pages. The graph also distinguishes by query length but the differences are diminishing. The mean Dice coefficient varies between 0.12 and 0.17. It is highest for the top two ranks and slightly decreases with higher ranks. The low mean Dice coefficients give an indication for a small degree of string similarity for the obtained results.

Table 3. Relative Retrieval Numbers for Titles, Lexical Signatures (LSs) and Tags, nDCG and MAP

	Top	**Top10**	**Top100**	**Undis**	**Mean nDCG**	**MAP**
Titles	**60.2**	4.2	0.6	**34.9**	**0.63**	**0.62**
LSs	36.5	6.6	1.3	55.6	0.4	0.39
Tags	22.1	**15.4**	**10.2**	52.4	0.32	0.27

4.3 Performance Compared to Content Based Queries

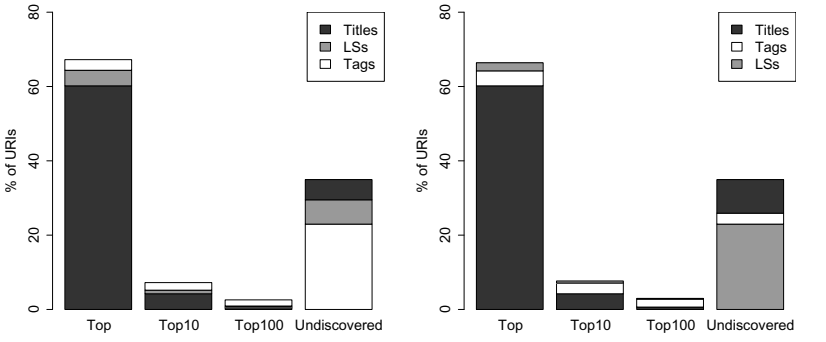
In order to give a comparison for the performance of tags we also apply two content based methods. We extract the title of each page and generate its lexical signature. We issue our three queries (title, lexical signature, tags) for each URI against the API. Table 3 summarizes their performance distinguished by our four retrieval cases, nDCG and MAP. Note that the data in Table 3 is based on aggregated values meaning we merged the results for 5− and 7−term lexical signatures into one category and likewise for all tag based query lengths. We can see that titles outperform lexical signatures, supporting our earlier findings in [9,10]. Both methods perform better than tags in terms of URIs returned top ranked, mean nDCG and MAP even though tags leave slightly fewer URIs undiscovered than lexical signatures. Tags return much more URIs in the top ten and top 100 than any other method. One interpretation of this observation is that tags, possibly rather generic by nature, are often not precise enough to return the URI top ranked. but they do provide enough specificity to return the pages within the top 100 results.

4.4 Combining Tags with Other Methods

Tables 2 and 3 show that the overall retrieval performance of tags alone is not very impressive. However what these tables do not show is the value of querying tags in combination with other methods. In other words, does the union of the results of more than one method improve the retrieval performance? And speaking from the preservation point of view, can we rediscover more missing URIs with combining two or even all three of the methods?

Extracting a web page's title from the content is cheap; it costs just one request to the resource. Lexical signatures are more expensive to generate since each term, as a candidate for the signature, requires the acquisition of a document frequency value. That means one request per unique term. Additionally we need to compute and normalize term frequency (TF) values. Obtaining tags, similar to titles, is very cheap because it only requires one request to Delicious.

With this cost model in mind we define two combinations of methods: *Title-Lexical_Signature-Tags* (*T-LS-TA*) and *Title-Tags-Lexical_Signature* (*T-TA-LS*). Since titles perform best (as seen earlier and also demonstrated in previous work [9]) we maintain their priority and query them first in both combinations. As our second step in *T-LS-TA* we apply the lexical signature based method to all URIs that remained undiscovered (34.9% as shown in Table 3). We thirdly



(a) Title-Lexical Signature-Tag Combination

(b) Title-Tag-Lexical Signature Combination

Fig. 2. Performance of Titles Combined with Lexical Signatures and Tags

apply the tag based method to all URIs that are still undiscovered in *T-LS-TA*. The difference in the second combination is that we apply the tag based method second (to the 34.9%) and the lexical signature based method third.

Figure 2 shows the combined retrieval performance. The data of combination *T-LS-TA* is shown in Figure 2(a) distinguished by contribution per method and separated in the previously introduced four retrieval categories. The first three bars (from left to right) are additive meaning the darkest part of the bars corresponds to the relative number of URIs returned by titles, the gray portion of the bars corresponds to the URIs not returned by titles but returned by lexical signatures and the white part of the bars represents the URIs neither returned by titles nor by lexical signatures. They are returned by tags only. Therefore these three left bars are to be read as if they were growing with the application of each additional method. The rightmost bar is to be read as if it was subtractive. For Figure 2(a) that means the dark portion of the bar represents the number of URIs undiscovered with titles (34.9%). The upper bound of the dark portion down to the upper bound of the gray portion represents the retrieval gain due to applying the second method. The height of the white portion of the bar corresponds to the final number of URIs that are left undiscovered after applying all three methods (23%) in the combination *T-LS-TA*. Figure 2(b) displays the data in the same way for the combination *T-TA-LS*. The color scheme remains the same with respect to the method meaning dark is still the title, gray still the lexical signature and white still represents tags.

The height of the gray bar for undiscovered URIs is of course identical to the corresponding white bar in Figure 2(a). The additive bar for the top ranked results is slightly higher in Figure 2(a) (67.2% vs. 66.4%) but the bars for the top ten and top 100 results are slightly higher in Figure 2(b) (7.2% vs. 7.7% and 2.6% vs. 3.0%). The results for the combination of methods in terms of

mean nDCG and MAP are summarized in Table 4. The performance increase of both combinations is statistically significant as determined by the t-test with p-values below the 0.05 threshold. Tags perform similarly compared to lexical signatures for URIs that remain undiscovered with the title method. Since tags are so much cheaper to obtain than lexical signatures these results lead to the recommendation to use tags a the default secondary method for rediscovering missing web pages in case tags are available through Delicious. This condition is crucial since we have seen that tags were rather sparse for previously analyzed web page corpora.

5 Ghost Tags

Previous research [4,5] has shown that about half the tags used to annotate URIs do not occur in the page's content. We find a slightly higher value with 66.3% of all tags not present in the page. If we consider the top ten tags only we find 51.5% of the tags not occurring in the page. This discrepancy intuitively makes sense since the ranking in Delicious is done by frequency of use which means that less frequently used tags are more likely to not appear in the page. However, these numbers only apply for the current version of the page. The tags provided by Delicious on the other hand are aggregated over an unknown period of time. The date of tags in Delicious can only be approximated but not reliably computed. It is possible that some tags used to occur in a previous version of the page and were removed or replaced at some point but still are available for that page through Delicious. We call these "ghost tags", terms that persist as tags after disappearing from the document itself.

To further investigate this aspect we use the Memento framework [15] to obtain old copies for all URIs that have tags not occurring in their content. For our dataset that applies to more than 95% of the URIs. Memento provides a timemap with references to all available Mementos (particular copies of a page at a certain point in time) per URI. Since we obtain different amounts of Mementos and different ages of the Mementos, we decided to only check tags against the first Memento meaning the oldest available copy of the page. We obtain Mementos of 3,306 URIs some of which date back to 1996. We find a total of 4.9% ghost tags. They occur in about one third of the previous versions of our web pages. Figure 3 displays the distribution of tags (dark gray) and the Mementos they occur in (light gray) per year. Note that the y-axis is plotted in log-scale. The vast majority of our "ghost tags" is found in Mementos from recent years especially in 2009. Only a few if any at all are found prior to 2006. We also see noticeable numbers

Table 4. Mean nDCG and Mean Average Precision for all Combinations of Methods

	TI	**TI-LS**	**TI-LS-TA**	**TI-TA**	**TI-TA-LS**
Mean nDCG	0.63	0.67	0.72	0.69	0.71
MAP	0.62	0.67	0.70	0.67	0.69

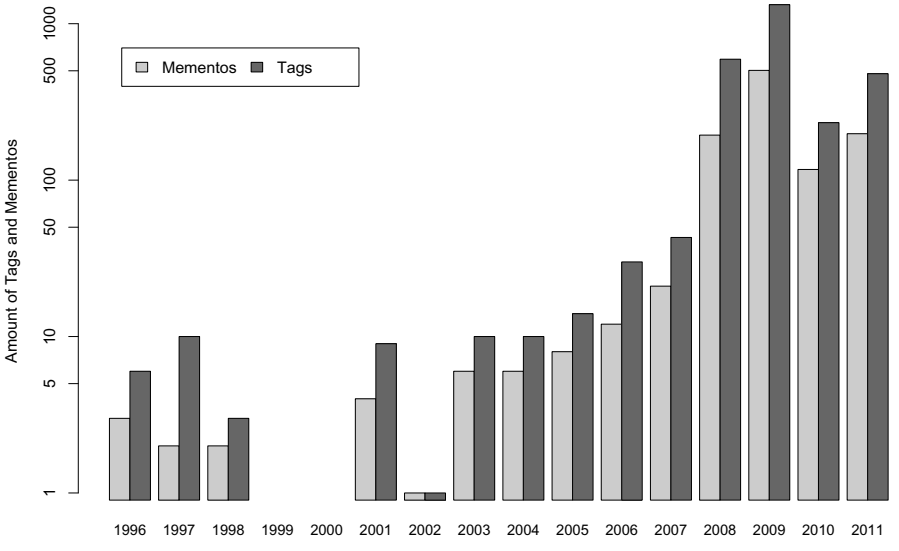


Fig. 3. Amount of Ghost Tags Occurring in Previous Versions of Web Pages

from 2011 which indicates a very short time between the publication of the tags at which time they did occur in the page and their disappearance from the page. The majority of these very recent Mementos were obtained from search engine caches. The observations from Figure 3 confirm: 1) ghost tags exist and better represent the past content of a web page than the current content, and 2) ghost tags are found in the more recent past and rarely date back more than three years.

We then determine the importance of ghost tags for a page. We compare the tags' occurrence frequency in Delicious and their term frequency (TF) in the first available Mementos. We rank each ghost tag according to its Delicious and its TF rank and normalize the rank to a value between zero and one in order to avoid a bias towards a greater amount of available tags and longer documents. The closer the value gets to zero the higher the rank and the greater the importance. Figure 4 displays the Delicious rank on the x-axis and the TF rank on the y-axis. Each dot represents one ghost tag. If a dot is plotted more than once, its shade gets darker (18 dots are plotted twice, one three times and one five times). The semi-transparent numbers indicate the percentage of dots or ghost tags in the corresponding quadrants. The numbers confirm our first visual impression of the graph. A majority of ghost tags (34.7%) occur in the first quadrant meaning their normalized Delicious rank is ≤ 0.5 and so is their TF rank. This indicates a high level of importance of the ghost tags for the document and also for the Delicious user. One fourth of the ghost tags

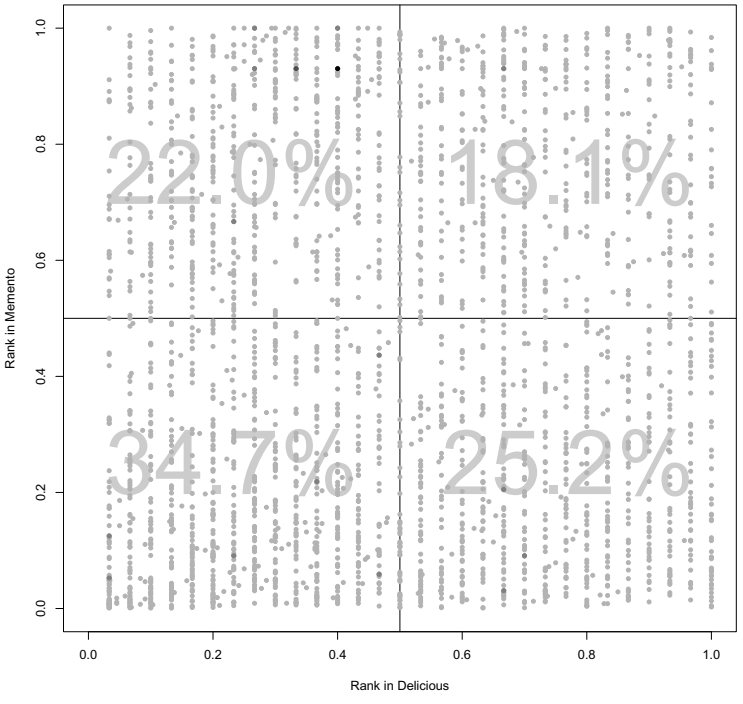


Fig. 4. Ghost Tags Ranks in Delicious and Corresponding Mementos

seem to be more important for the document than in Delicious since their ranking there is > 0.5. On the other hand for 22% of ghost tags the inverse holds true. In 18.1% of the cases we can claim that "only" infrequent terms became ghost tags. These results show the significance of ghost tags since one third of them were used very frequently in the document and still are used frequently in Delicious.

6 Conclusions and Future Work

In this paper we have investigated the performance of tags for the purpose of discovering missing web pages. We obtained tags of almost $5,000$ URIs from Delicious and showed that a search engine query containing five to eight tags performs best. More than 20% of the URIs are returned in the top ten ranks. We have further provided evidence for the top ten results to be similar to the URI the queried tags were obtained from. Compared to querying the title of the page or its lexical signature tags do not perform well but a combination of these methods increases the overall retrieval performance. We have also explored the notion of "ghost tags" as terms from Delicious that do not occur in the current

version but do occur in a previous version of the web page. More than one out of three ghost tags appear to be important for the user as well as for the document since they rank high in Delicious and occur frequently in the text.

Our notion of ghost tags refers to the earliest available copy of web pages only. We will further investigate the aspect of time by including more copies of pages over time giving us a more precise idea of the age of the ghost tags. Another unanswered question is whether some tags predate the actual web page. If we can timestamp tags and monitor their frequency of use we can give a more specific description of their dynamics. In other words do users stop using tags when they disappear from or appear in the page? We have shown in previous work that titles and lexical signatures of web pages change over time. Naturally these methods after some time become obsolete as search engine queries. The question remains whether tags, as user given keywords, must be seen as dated at some point as well.

References

1. Agichtein, E., Zheng, Z.: Identifying "Best Bet" Web Search Results by Mining Past User Behavior. In: Proceedings of KDD 2006, pp. 902–908 (2006)
2. Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., Su, Z.: Optimizing Web Search Using Social Annotations. In: Proceedings of WWW 2007, pp. 501–510 (2007)
3. Berners-Lee, T.: Cool URIs don't change (1998), http://www.w3.org/Provider/Style/URI.html
4. Bischoff, K., Firan, C., Nejdl, W., Paiu, R.: Can All Tags Be Used for Search? In: Proceedings of CIKM 2008, pp. 193–202 (2008)
5. Heymann, P., Koutrika, G., Garcia-Molina, H.: Can Social Bookmarking Improve Web Search? In: Proceedings of WSDM 2008, pp. 195–206 (2008)
6. Jason Morrison, P.: Tagging and Searching: Search Retrieval Effectiveness of Folksonomies on the World Wide Web. Information Processing and Management 44, 1562–1579 (2008)
7. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately Interpreting Clickthrough Data as Implicit Feedback. In: Proceedings of SIGIR 2005, pp. 154–161 (2005)
8. Klein, M., Nelson, M.L.: Revisiting lexical signatures to (Re-)Discover web pages. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 371–382. Springer, Heidelberg (2008)
9. Klein, M., Nelson, M.L.: Evaluating Methods to Rediscover Missing Web Pages from the Web Infrastructure. In: Proceedings of JCDL 2010, pp. 59–68 (2010)
10. Klein, M., Shipman, J., Nelson, M.L.: Is This a Good title? In: Proceedings of Hypertext 2010, pp. 3–12 (2010)
11. Klein, M., Ware, J., Nelson, M.L.: Rediscovering Missing Web Pages Using Link Neighborhood Lexical Signatures. In: Proceedings of JCDL 2011 (2011)
12. Krause, B., Hotho, A., Stumme, G.: A Comparison of Social Bookmarking with Traditional Search. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 101–113. Springer, Heidelberg (2008)

13. Marshall, C.C., McCown, F., Nelson, M.L.: Evaluating personal archiving strategies for internet-based information (2007)
14. Sugiyama, K., Hatano, K., Yoshikawa, M., Uemura, S.: Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In: Proceedings of HYPERTEXT 2003, pp. 198–207 (2003)
15. Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L., Ainsworth, S., Shankar, H.: Memento: Time Travel for the Web. Technical Report arXiv:0911.1112 (2009)
16. Yanbe, Y., Jatowt, A., Nakamura, S., Tanaka, K.: Can Social Bookmarking Enhance Search in the Web? In: Proceedings of JCDL 2007, pp. 107–116 (2007)

Mapping MPEG-7 to CIDOC/CRM

Anastasia Angelopoulou, Chrisa Tsinaraki, and Stavros Christodoulakis

TUC/MUSIC, Technical University of Crete, University Campus, Kounoupidiana, 73100 Chania, Crete, Greece
{anastasia,chrisa,stavros}@ced.tuc.gr

Abstract. The MPEG-7 is the dominant standard for multimedia content description; thus, the audiovisual Digital Library contents should be described in terms of MPEG-7. Since there exists a huge amount of audiovisual content in the cultural heritage domain, it is expected that several cultural heritage objects, as well as entities related with them (i.e. people, places, events etc.), have been described using MPEG-7. On the other hand, the dominant standard in the cultural heritage domain is the CIDOC/CRM; consequently, the MPEG-7 descriptions cannot be directly integrated in the cultural heritage digital libraries.

We present in this paper a mapping model and a system that allow the transformation of the MPEG-7 descriptions to CIDOC/CRM descriptions, thus allowing the exploitation of multimedia content annotations in the cultural heritage digital libraries. In addition, the proposed mapping model allows linking MPEG-7 descriptions to CIDOC/CRM descriptions in a Linked Data scenario.

Keywords: MPEG-7, CIDOC/CRM, Mapping, Multimedia, Cultural Heritage.

1 Introduction

The MPEG-7 is the dominant standard for multimedia content description [1] and allows describing all the aspects of the multimedia content, including the semantics, the low-level image, audio and motion features, the structural information, the media-related information etc. The MPEG-7 descriptions can be captured automatically (using, for example, camera sensors) or semi-automatically. The amount of multimedia data captured daily is increasing extremely fast due to the proliferation of inexpensive cameras and associated sensors (see for example [17]). Very frequently the digital audiovisual content captured contains cultural heritage objects.

On the other hand, the dominant standard in the cultural heritage domain for the description of cultural heritage objects (i.e museum exhibits, archival material, books etc.) is the CIDOC/CRM [2]. Several other standards used in the cultural heritage domain have been mapped to the CIDOC/CRM (like [3] [4]). However, the MPEG-7 descriptions of the audiovisual content cannot be directly integrated in the cultural heritage digital libraries since the MPEG-7 has not been mapped to the CIDOC/CRM.

The need for transforming MPEG-7 descriptions to CIDOC/CRM ones as well as the capability of linking them in CIDOC/CRM descriptions in a Linked Data scenario has been recognized by important consortiums in the Digital Library domain like, for example, the Europeana consortium [5]. Europeana aims to develop a European

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 40–51, 2011.

digital library containing digitized material about the European scientific and cultural heritage. In particular, the *Europeana Data Model (EDM)* [7] has adopted the CIDOC/CRM core, while the consortium emphasizes the need for linking existing descriptions of the digitized material in the EDM descriptions [6], according to the linked data approach [8].

The previous research in interoperability support between MPEG-7 and CIDOC/CRM focuses on the representation of CIDOC/CRM descriptions in MPEG-7 syntax. In particular, [9] proposes specific extensions to the CIDOC/CRM in order to be able to accommodate the temporal and spatial aspects of information objects, while [10] has developed a methodology that allows automatically generating semantic MPEG-7 multimedia annotations from CIDOC/CRM descriptions. However, in both cases, the inverse functionality, which should allow the transformation and/or linking of MPEG-7 multimedia annotations in CIDOC/CRM descriptions, is missing.

We present in this paper *MPEG72CIDOC*, a *mapping model* that maps the MPEG-7 constructs to CIDOC/CRM constructs and a software component that, based on MPEG72CIDOC, allows the transformation of MPEG-7 descriptions to CIDOC/CRM descriptions as well as linking them to CIDOC/CRM descriptions. These mechanisms allow the multimedia content descriptions to be exploited in the cultural heritage digital libraries. This work complements our previous research for the transformation of CIDOC/CRM descriptions in MPEG-7 syntax [10]. The MPEG72CIDOC mapping model differs from that of [10] in the following: *(a)* It has adopted the MPEG-7 viewpoint for mapping the MPEG-7 constructs to CIDOC/CRM constructs. As a consequence, there do not exist corresponding CIDOC/CRM constructs for some MPEG-7 constructs (like, for example, the spatial relations above, south, left etc.) and some of the mappings specified in [10] for the inverse process are not appropriate from this viewpoint, since the two standards describe several aspects in different levels of granularity; and *(b)* It takes into account all the MPEG-7 MDS and not only the semantic part, as was done in [10]. Moreover, the implementation of the MPEG72CIDOC mapping model in order to allow the automatic transformation of MPEG-7 descriptions to CIDOC/CRM ones has been integrated in the toolkit developed in [10] for the automatic transformation of CIDOC/CRM descriptions in MPEG-7 syntax.

The MPEG72CIDOC mappings may also be applied between CIDOC/CRM and any ontology like [12] that captures the MPEG-7 semantics, since the mappings have not been based on the XML Schema syntax of MPEG-7. They also allow the exploitation of user preferences that have been expressed for multimedia content described using MPEG-7 [14] in the cultural heritage domain.

In addition, the transformation of the MPEG-7 descriptions to CIDOC/CRM ones allows using the Semantic Web technologies over the transformed descriptions in the cultural heritage domain, without having to use any MPEG-7 based ontology and thus not having to face the interoperability issues arising from the existence of several MPEG-7 based ontologies [13].

The rest of this paper is structured as follows: The MPEG72CIDOC mapping model is presented in Section 2, the MPEG-7 to CIDOC/CRM transformation process is described in Section 3, the implementation is discussed in Section 4, a transformation example is presented in Section 5 and the paper concludes in Section 6, which also outlines our future research directions.

2 The MPEG72CIDOC Mapping Model

In this section we present the MPEG72CIDOC mapping model that we have developed in order to allow the exploitation of MPEG-7 descriptions in CIDOC/CRM working environments.

The MPEG-7 focuses on multimedia content description, while the CIDOC/CRM focuses on cultural heritage concepts. Thus, the MPEG-7 provides a more extended set of description tools for multimedia content description, while the CIDOC/CRM provides a fine-grained conceptualization within the cultural heritage domain. As a consequence, we faced the problem of the accurate representation of the multimedia-specific MPEG-7 concepts in CIDOC/CRM. This problem was solved through the representation of these concepts using the CIDOC/CRM entity "E55.Type", which is an extensibility mechanism of the CIDOC/CRM model and its instances can be considered as classes that are organized in class hierarchies using the properties "P127 has broader term/has narrower term". The association of an "E55.Type" entity instance to its type is implemented through the property "P2 has type".

In the MPEG72CIDOC mapping model, the MPEG-7 *types*, which represent the MPEG-7 concepts, are mapped to semantically correspondent CIDOC/CRM *entities*. Moreover, the MPEG-7 *relations*, which associate instances of the MPEG-7 types, are mapped to CIDOC/CRM *properties* that associate CIDOC/CRM entities.

There are two types of mappings in the MPEG72CIDOC mapping model: a) *static* mappings, which essentially are *correspondences* [15] between the MPEG-7 and the CIDOC/CRM constructs and are specified at design-time; and b) *conditional* mappings that are evaluated in real-time according to the given context since they are based on *mapping rules* [16] that have been specified at design-time.

MPEG-7 Type Mappings. The MPEG72CIDOC mapping model is based on the following principles for mapping the MPEG-7 types to CIDOC/CRM entities:

- For every MPEG-7 type *mt* that can be directly mapped to a CIDOC/CRM entity *ce*, an exact static mapping between *mt* and *ce* is defined. For example, the MPEG-7 type "PersonType" that represents persons is mapped to the semantically correspondent CIDOC/CRM entity "E21 Person".
- Every MPEG-7 type *mmt* that represents a multimedia-specific concept for which does not exist a corresponding CIDOC/CRM entity is mapped to an instance *ramd* of the CIDOC/CRM entity "E55 Type". For example, the MPEG-7 type "VideoType" is mapped to the "VideoType" instance of the CIDOC/CRM entity "E55 Type".
- The MPEG-7 provides *abstraction* support, which allows the representation of both instance-level semantic abstract descriptions and class-level semantic descriptions. In particular, the representation of the abstract MPEG-7 description *amd* is based on the value of its *dimension* attribute that indicates its abstraction level:
  - If *dimension* has a value greater than or equal to 1, *amd* is a class-level abstract semantic description that represents a class and is mapped to an instance *tme* of the CIDOC/CRM entity "E55 Type". For example, an abstract MPEG-7 description that represents a class of buildings is mapped to an instance of the CIDOC/CRM entity "E55 Type".

  - If *dimension* has a value of 0, *amd* is an instance-level semantic description independent from the multimedia content and it describes a reusable instance (e.g. Parthenon). In this case *amd* is a concrete semantic description and is represented by an instance *tme* of the CIDOC/CRM entity "E77 Persistent Item". For example, an abstract MPEG-7 description that represents Parthenon is mapped to an instance of the CIDOC/CRM entity "E77 Persistent Item".
- The representation *rte* of any element *te* of the MPEG-7 type *mt* is associated with the representation *rmt* of *mt* using one of the following CIDOC/CRM properties:
  - **P141 assigned** if *te* is an object.
  - **P140 assigned attribute to** if *te* is a relation.
- The MPEG-7 attributes "id", "href", "xml:lang" and "xsi:type" are transformed to the appropriate CIDOC/CRM properties using specialized algorithms that are described in Section 3.

Due to the large number of the MPEG-7 types, the presentation of the MPEG72CIDOC mappings here is non-exhaustive (an exhaustive presentation is available in [11]). An excerpt of the MPEG72CIDOC mappings between MPEG-7 types and CIDOC/CRM entities is shown in Table 1.

Table 1. Excerpt of the MPEG72CIDOC MPEG-7 Type Mappings

MPEG-7 Type	CIDOC/CRM Entity
VideoType	E55 Type ("VideoType")
AgentType	E39 Actor
MultimediaContentType	E31 Document
PersonType	E21 Person
…	…

MPEG-7 Relation Mappings. The MPEG72CIDOC mapping of an MPEG-7 relation *mr* to a CIDOC/CRM property *cp* falls in one of the following categories:

- **Exact mapping.** In this case, the MPEG-7 relation *mr* is mapped to the CIDOC/CRM property *cp* that has exactly the same meaning. For example, the MPEG-7 relation "inside" is mapped to the semantically correspondent CIDOC/CRM property "P89 falls within".
- **Mapping to the closest meaning.** In this case *mr* is mapped to the CIDOC/CRM property *cp* with the closest semantic meaning. For example, the MPEG-7 relation "key" is mapped to the CIDOC/CRM property "P1 is identified by".
- **No Mapping.** In this case *mr* cannot be mapped to a CIDOC/CRM property, since there does not exist a CIDOC/CRM property with the same (or at least similar) semantics. For example, the MPEG-7 relation "above" is not mapped to any CIDOC/CRM property.
- **Conditional Mapping.** In this case, *mr* is mapped to different CIDOC/CRM properties based on the type of its source *mrs* and its target *mrt*. This happens if *mr* is of type *location*, *location of* or *overlaps*:

- If *mr* is of *location* type, then: *(a)* If *mrs* is an event, *mr* is mapped to the CIDOC/CRM property "P7 took place at"; and *(b)* If *mrs* is an object, *mr* is mapped to the CIDOC/CRM property "P53 has former or current location".
- If *mr* is of *location of* type, then: *(a)* If *mrt* is an event, *mr* is mapped to the CIDOC/CRM property "P7 witnessed"; and *(b)* If *mrt* is an object, *mr* is mapped to the CIDOC/CRM property "P53 is former or current location of".
- If *mr* is of *overlaps* type, then: *(a)* If *mrs* is a place, *mr* is mapped to the CIDOC/CRM property "P121 overlaps with"; and *(b)* If *mrs* is a time period, *mr* is mapped to the CIDOC/CRM property "P132 overlaps with".

An excerpt of the MPEG72CIDOC MPEG-7 relation mappings is shown in Table 2 (an exhaustive presentation of the mappings is available in [11]).

Table 2. Excerpt of the MPEG72CIDOC MPEG-7 Relation Mappings

MPEG-7 Relation	CIDOC/CRM property
Exact mapping	
inside	P89 falls within
precedes	P120 occurs before
agent	P14 carried out by
depictedBy	P62 is depicted by
…	…
Mapping to the closest meaning	
refines	P70 documents
user	P125 used object of type
key	P1 is identified by
goal	P21 had general purpose
…	…
No Mapping	
south	-
left	-
above	-
…	…

3 MPEG-7 to CIDOC/CRM Transformation

In this section we present the MPEG-7 to CIDOC/CRM transformation process that implements the MPEG72CIDOC mapping model.

The MPEG-7 to CIDOC/CRM transformation process is outlined in Fig. 1. The transformation of an MPEG-7 description *md* starts by locating all the elements of *md*. Then the MPEG-7 relation elements are separated, the transformation of the *md* elements and relations takes place and the produced CIDOC/CRM description is finalized after the association of the transformations of the individual MPEG-7 constructs (i.e. elements and relationships).

The transformation of the MPEG-7 elements is outlined in Fig. 2 (due to space limitations, details on the MPEG-7 element transformation are available in [11]): For every MPEG-7 element *e* the element name and value are located first. Then the

mapping of the type of *e* to the appropriate CIDOC/CRM entity *ce* is used for the transformation of *e* in an instance *re* of *ce*. At this stage it is also checked if *e* has attributes. If this is the case, they are transformed in CIDOC/CRM properties. Finally, the representation *re* of *e* is associated with the CIDOC/CRM properties that represent its attributes and is added in the CIDOC/CRM description.

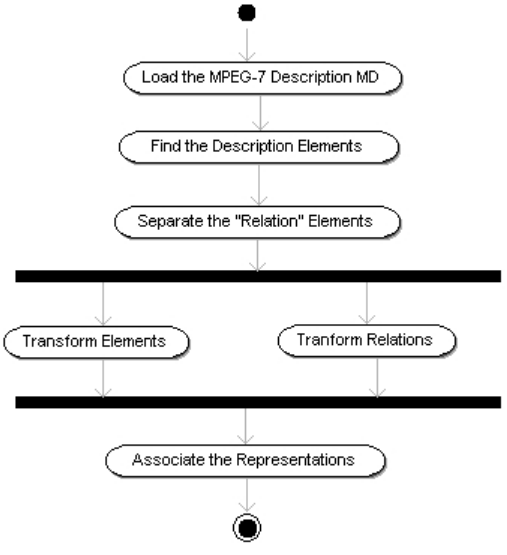


Fig. 1. The overall MPEG-7 to CIDOC/CRM transformation process

During the transformation of a "Relation" element *mre* the source, the type and the target of *mre* are located. The source and the target of *mre* are transformed, respectively, in the CIDOC/CRM entities *crs* and *crt* and the type of *mre* is transformed in a CIDOC/CRM property *cp* that has *crs* as range and *crt* as domain.

The transformation of the MPEG-7 element attributes depends on the attribute type. In particular, the following hold for the transformation of an attribute *a* of the element *e*, where *e* has been transformed to an instance *re* of the CIDOC/CRM entity *ce* (details on the MPEG-7 attribute transformation are available in [11]):

- If *a* is the "xsi:type" attribute, a CIDOC/CRM individual *at* of type "E55 Type" is created and *a* is transformed in the CIDOC/CRM property "P2 has type" that associates *at* with *re*.
- If *a* is the "id" attribute, a CIDOC/CRM individual *ai* of type "E42 Identifier" is created and *a* is transformed in the CIDOC/CRM property "P1 is identified by" that associates *ai* with *re*.
- If *a* is the "xml:lang" attribute, a CIDOC/CRM individual *al* of type "E56 Language" is created and *a* is transformed in the CIDOC/CRM property "P72 has language" that associates *al* with *re*.

– If *a* is the "href" attribute, a CIDOC/CRM individual *aio* of type "E73 information Object" is created and *a* is transformed in the CIDOC/CRM property "P67 refers to" that associates *aio* with *re*.

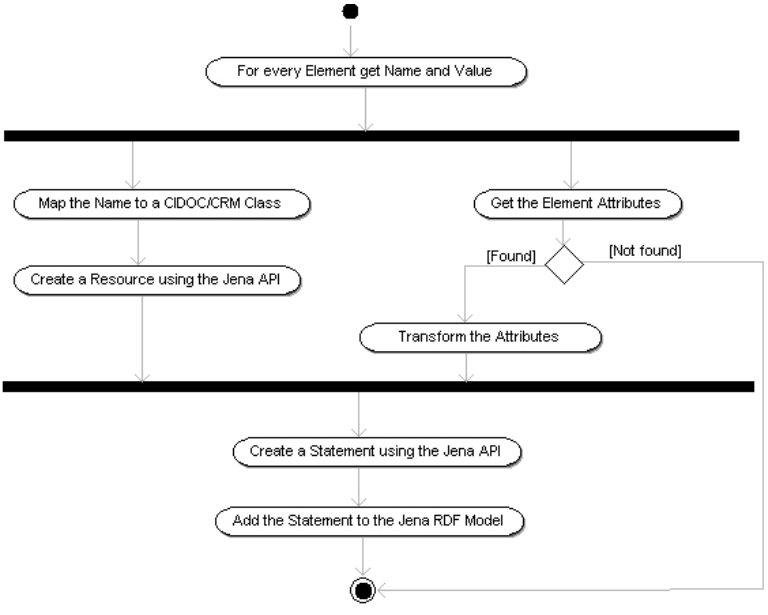


Fig. 2. MPEG-7 Element Transformation

4 Implementation

In this section we present the software that implements the MPEG-7 to CIDOC/CRM transformation process described in section 3. This software extends with the MPEG-7 to CIDOC/CRM transformation functionality the transformation toolkit developed in [10], which allows the automatic transformation of CIDOC/CRM descriptions to valid MPEG-7 multimedia object descriptions.

The toolkit provides a Graphical User Interface that allows the user to see a graphical representation of loaded and generated descriptions (MPEG-7 and CIDOC/CRM descriptions). A screenshot of the toolkit is presented in Fig. 3.

The toolkit GUI is divided in two panels: the *function panel* on the left, and the *mapping panel* on the right. The function panel contains all the necessary buttons for the user actions, such as loading descriptions, saving the generated documents, performing conversions between CIDOC/CRM and MPEG-7 descriptions, and presenting the graphs of the loaded and generated descriptions. The mapping panel shows an MPEG-7 description on the left side and the equivalent CIDOC/CRM description on the right side.

The toolkit has been implemented using the Java programming language, the XML Beans framework [6] for the manipulation of the MPEG-7 XML documents and the Jena framework [5] for parsing the CIDOC/CRM descriptions (in RDF syntax).

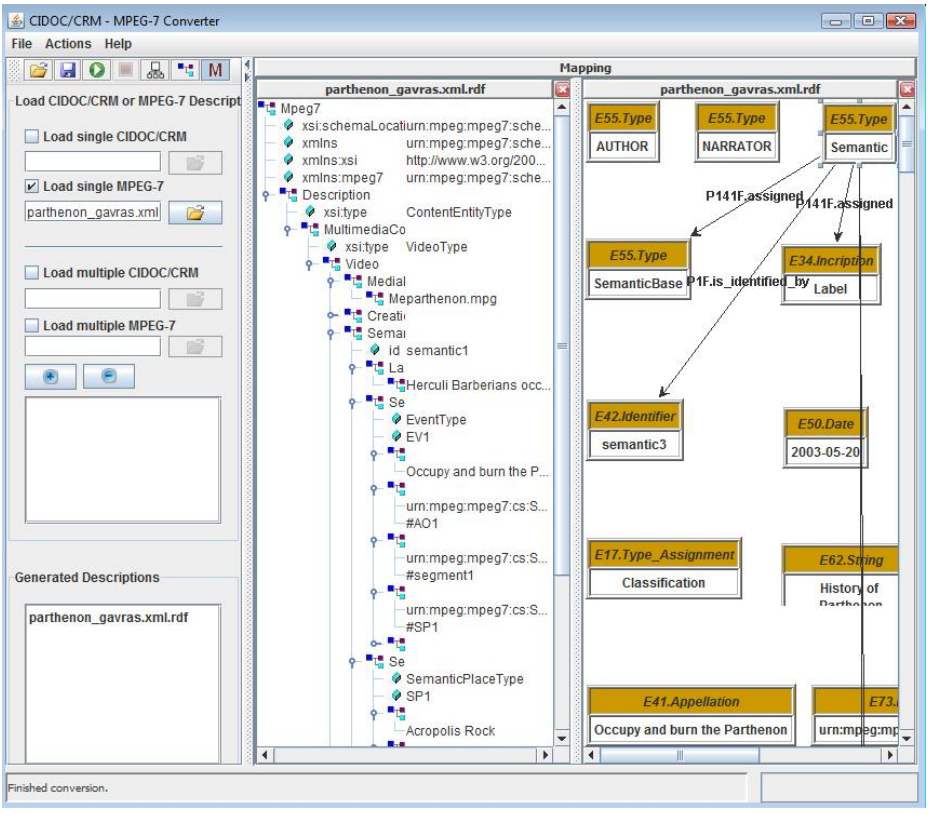


Fig. 3. Screenshot of the toolkit

5 Transformation Example

We provide in this section a short example of the MPEG-7 to CIDOC/CRM transformation, in order to demonstrate the entire transformation methodology and understand how the MPEG-7 to CIDOC/CRM transformation is applied on a real-world description. In particular, we used a part of the MPEG-7 description "Parthenon by Costas Gavras" (Fig. 4), which describes the event "Occupy and burn the Parthenon" in a video on the history of Parthenon.

```
<Description xsi:type="ContentEntityType">
 <MultimediaContent xsi:type="VideoType">
  <Video id="parthenon">
   <MediaLocator>
    <MediaUri>parthenon.mpg</MediaUri>
   </MediaLocator>
   <CreationInformation>
    <Creation>
     <Title xml:lang="en">Parthenon</Title>
    <Creator>
      <Role href="urn:mpeg:mpeg7:cs:RoleCS:2001:AUTHOR"/>
```

Fig. 4. An excerpt of the MPEG-7 description of the video «Parthenon by Costas Gavras»

```
    <Agent xsi:type="PersonType">
     <Name>
      <GivenName>Costas</GivenName>
      <FamilyName>Gavras</FamilyName>
     </Name>
    </Agent>
   </Creator>
   <CopyrightString>Hellenic Culture Organisation S.A.</CopyrightString>
  </Creation>
 </CreationInformation>
 <Semantic id="semantic1">
  <Label>
   <Name>Herculi Barberians occupy and burn the Parthenon </Name>
  </Label>
  <SemanticBase xsi:type="EventType" id="EV1">
   <Label>
    <Name> Occupy and burn the Parthenon </Name>
   </Label>
   <Relation target="#AO1" type="agent"/>
   <Relation target="#segment1" type="depictedBy"/>
   <Relation target="#SP1" type="location"/>
   <Relation target="#ST1" type=" time"/>
  </SemanticBase>
  <SemanticBase xsi:type="SemanticPlaceType" id="SP1">
   <Label>
    <Name> Acropolis Rock </Name>
   </Label>
   <Place>
    <Name xml:lang="en">Acropolis Rock in the City of Athens</Name>
    <Region> gr </Region>
   </Place>
  </SemanticBase>
  <SemanticBase xsi:type="SemanticTimeType" id="ST1">
   <Label><Name> 267 A.D.</Name></Label>
   <Relation source="#ST1" target="#ST2" type="precedes"/>
  </SemanticBase>
  <SemanticBase xsi:type="AgentObjectType" id="AO1">
   <Label><Name>Herculi Barberians</Name></Label>
   <Agent xsi:type="OrganizationType">
    <Name>Herculi Barberians</Name>
   </Agent>
  </SemanticBase>
 </Semantic>
 <MediaTime>
  <MediaTimePoint>T00:00:00</MediaTimePoint>
  <MediaDuration>PT07M33S</MediaDuration>
 </MediaTime>
 <TemporalDecomposition gap="false" overlap="false">
  <VideoSegment id="segment1">
   <TextAnnotation>
    <FreeTextAnnotation>
     267 A.D. Herculi Barberians occupy and burn the Parthenon
    </FreeTextAnnotation>
   </TextAnnotation>
   <Relation target="key1.gif" type="key"/>
   <Relation target="segment1.rm" type="representedBy"/>
   <MediaTime>
    <MediaTimePoint>T00:01:22</MediaTimePoint>
    <MediaDuration>PT00M09S</MediaDuration>
   </MediaTime>
```

Fig. 4. (*continued*)

```
    </VideoSegment>
   </TemporalDecomposition>
  </Video>
 </MultimediaContent>
</Description>
```

Fig. 4. (*continued*)

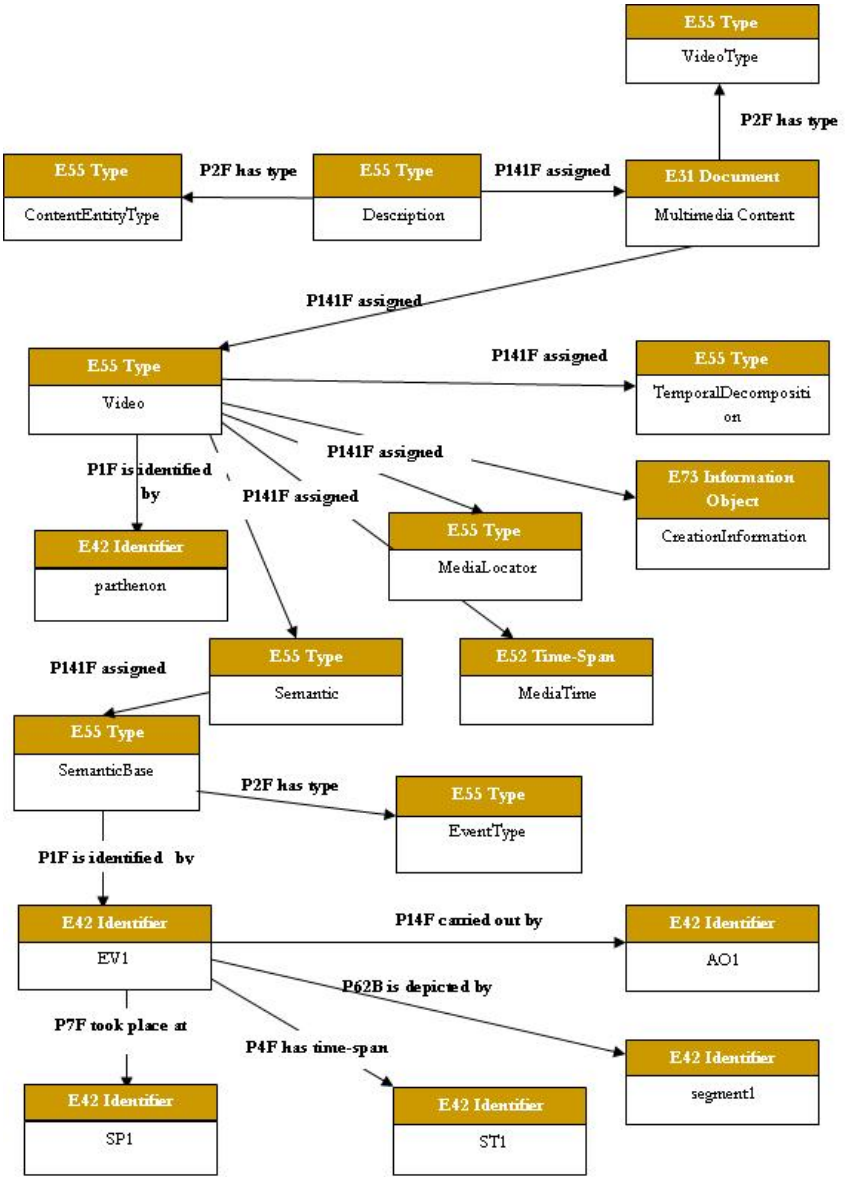


Fig. 5. An excerpt of the CIDOC/CRM description generated for the (the bold regions of the) MPEG-7 description of Fig. 1

When the transformation process starts, the MPEG-7 constructs are located and classified in the following categories:

- **Elements:** "Description", "MultimediaContent", "Video", "MediaLocator", "MediaUri", "CreationInformation", "Creation", "Title", "Abstract", "Creator", "FreeTextAnnotation", "Role", "Agent", "Name", "FamilyName", "GivenName", "CopyrightString", "Region", "Semantic", "Label", "SemanticBase", "Place", "MediaTime", "MediaTimePoint", "MediaDuration", "TemporalDecomposition", "VideoSegment", "TextAnnotation".
- **Relations:** "agent", "depictedBy", "location", "time", "key", "representedBy", "precedes".

According to the activity diagram of Fig. 1, a different set of steps is followed for the constructs that belong to each category. An excerpt of the CIDOC/CRM description generated for the MPEG-7 description of Fig. 4 is shown in Fig. 5.

Notice that the MPEG-7 elements have been transformed to instances of the CIDOC/CRM entities that have been mapped to their element types. For example, recall that the MPEG-7 type "MultimediaContentType" has been mapped to the CIDOC/CRM entity "E31 Document", and notice that the MPEG-7 element "MultimediaContent", of type "MultimediaContentType", has been transformed to an instance of "E31 Document".

Notice also that the MPEG-7 relations have been transformed to the mapped CIDOC/CRM properties. For example, recall that the MPEG-7 relation "depictedBy" has been mapped to the CIDOC/CRM property "P62 is depicted by" and notice that the MPEG-7 relation "depictedBy" has been transformed to the CIDOC/CRM property "P62 is depicted by".

6 Conclusions – Future Work

In this paper we have presented the MPEG72CIDOC mapping model, which allows the transformation of MPEG-7 descriptions to CIDOC/CRM descriptions as well as linking them to CIDOC/CRM descriptions, and a software that implements it. Using our methodology and software the multimedia content annotations can be exploited in the cultural heritage digital libraries. This work complements our previous research for the transformation of CIDOC/CRM descriptions in MPEG-7 syntax [10].

Since the EDM has adopted the CIDOC/CRM core, the work presented here is a first step towards supporting the transformation and/or linking of MPEG-7 descriptions to EDM descriptions in a Linked Data scenario.

Our future research includes: *(a)* The definition of a two-way mapping between the MPEG-7 and the EDM, which will allow full interoperability support among these standards. Such functionality is very important for the Digital Library community; and *(b)* The extensive evaluation of the MPEG72CIDOC mapping model over real-world datasets.

Acknowledgment. This work was partially supported by the FP7 project Natural Europe (Project Ref. No 250579, Area CIP-ICT-PSP.2009.2.5 - Digital Libraries).

References

1. Salembier, P.: MPEG-7 Multimedia Description Schemes. IEEE TA on Circuits and Systems for Video Technology 11(6), 748–759 (2001)
2. ISO 21127:2006 Information and documentation – A reference ontology for the interchange of cultural heritage information (CIDOC/CRM)
3. Gaitanou, P., Gergatsoulis, M.: Mapping VRA Core 4.0 to the CIDOC CRM ontology. In: The Proc. of the 1st Workshop on Digital Information Management, Corfu, Greece, March 30-31 (2011)
4. Doerr, M., Leboeuf, P.: Modelling intellectual processes: The FRBR—CRM harmonization. In: Conference Proceedings of ICOM-CIDOC Annual Meeting, Gothenburg, Sweden, pp. 10–14 (2006)
5. The Europeana consortium, `http://www.europeana.eu/`
6. Zeinstra, M., Keller, P.: Open Linked Data and Europeana. In: Europeana Business Documents (2010), `http://version1.europeana.eu/c/document_library/get_file?uuid=374c381f-a48b-4cf0-bbde-172cf03672a2&groupId=10602`
7. Isaac, A. (ed.): Europeana Data Model Primer. Europeana technical documents, `http://version1.europeana.eu/c/document_library/get_file?uuid=718a3828-6468-4e94-a9e7-7945c55eec65&groupId=10605`
8. WWW Consortium (W3C). Open Linked Data, `http://www.w3.org/standards/semanticweb/data`
9. Hunter, J.: Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums. In: The Proc. of the Museums and the Web International Conference, Boston (April 2002)
10. Ntousias, A., Gioldasis, N., Tsinaraki, C., Christodoulakis, S.: Rich Metadata and Context Capturing through CIDOC/CRM and MPEG-7 Interoperability. In: The Proc. of the ACM Conference on Image and Video Retrieval (ACM CIVR), pp. 151–160 (2008)
11. Angelopoulou, A., Tsinaraki, C., Christodoulakis, S.: Mapping MPEG-7 to CIDOC/CRM. Technical report, TUC/MUSIC (2011), `http://www.music.tuc.gr/GetFile?FILE_TYPE=PUB.FILE&FILE_ID=359`
12. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability support between MPEG-7/21 and OWL in DS-MIRF. IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE), Special Issue on the Semantic Web Era 19(2), 219–232 (2007)
13. Troncy, R., Celma, O., Little, S., Garcia, R., Tsinaraki, C.: MPEG-7 based Multimedia Ontologies: Interoperability Support or Interoperability Issue? In: The Proc. of the MaReSO Workshop (December 5, 2007)
14. Tsinaraki, C., Christodoulakis, S.: A Multimedia User Preference Model that Supports Semantics and its Application to MPEG 7/21. In: The Proc. of the IEEE Multimedia Modeling 2006 Conference (IEEE MMM 2006), Beijing, China, pp. 35–42 (January 2006)
15. Euzenat, J., Shvaiko, P.: Ontology matching, p. 42. Springer, Heidelberg (DE) (2007)
16. Euzenat, J., Shvaiko, P.: Ontology matching, p. 43. Springer, Heidelberg (DE) (2007)
17. Christodoulakis, S., Foukarakis, M., Ragia, L., Uchiyama, H., Imai, T.: Picture Context Capturing for Mobile Databases. IEEE MultiMedia 17(2), 34–41 (2010)

A Language Independent Approach for Named Entity Recognition in Subject Headings

Nuno Freire[1,2], José Borbinha[1], and Pável Calado[1]

[1] Instituto Superior Técnico, Technical University of Lisbon,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2] The European Library, National Library of the Netherlands,
Willem-Alexanderhof 5, 2509 LK The Hague, Netherlands
{nuno.freire,jlb,pavel.calado}@ist.utl.pt

Abstract. Subject headings systems are tools for organization of knowledge that have been developed over the years by libraries. The SKOS Simple Knowledge Organization System has provided a practical way to represent subject headings systems using the Resource Description Framework, and several libraries have taken the initiative to make subject headings systems widely available as open linked data. Each individual subject heading describes a concept, however, in the majority of cases, one subject heading is actually a combination of several concepts, such as a topic bounded in geographical and temporal scopes. In these cases, the label of the concept actually carries several concepts which are not represented in structured form. Our work explores machine learning techniques to recognize the sub concepts represented in the labels of SKOS subject headings. This paper describes a language independent named entity recognition technique based on conditional random fields, a machine learning algorithm for sequence labelling. This technique was evaluated on a subset of the Library of Congress Subject Headings, where we measured the recognition of geographic concepts, topics, time periods and historical periods. Our technique achieved an overall F_1 score of 0.98.

Keywords: named entity recognition, subject headings, linked data, SKOS, machine learning.

1 Introduction

Subject headings systems are tools for organization of knowledge, which have been developed over the years by libraries. Assignment of subject headings to the items within their collections is a part of bibliographic organization tasks carried out by libraries. Subject headings aid the user to discover items in the catalogue that pertain to similar subject matter [1].

Subject headings systems, like other knowledge organization systems such as thesauri and taxonomies, can nowadays be more widely used if made available within the framework of the Semantic Web. The SKOS Simple Knowledge Organization System[1] [2] has

[1] http://www.w3.org/2004/02/skos/

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 52–61, 2011.

been developed for this purpose, and it provides a practical way to represent subject headings systems using the Resource Description Framework.

Several libraries have taken the initiative to make subject headings systems widely available by representing them in SKOS, and making them available as open linked data. Some known examples are the Library of Congress Subject Headings[2] (LCSH), the *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*[3] (RAMEAU), and *Schlagwortnormdatei*[4] (SWD), which are subject headings systems in English, French and German, respectively.

Each individual subject heading describes a concept. However, in the majority of cases, one subject heading is actually a combination of several concepts, such as a topic bounded in geographical and temporal scopes. Although the concept is available in SKOS, and therefore available with some semantics for machine processing, its individual subtopics are not, which limits what machines can inference from the subject headings.

As subject heading systems become available as open linked data, the value of linking all these sub concepts to their representation in other open data sets becomes more relevant. Several millions of resources have assigned subject headings, in libraries catalogues and digital libraries. Improving the semantics of subject headings has the potential to benefit the retrieval and access to all these resources.

In our work, we explored machine learning techniques to recognize the sub concepts in subject headings. This paper describes a named entity recognition technique developed for the particular case of subject headings. It is based on conditional random fields [4], a machine learning algorithm for sequence labelling, and was designed to be language independent so that it can be applied to the many subject headings systems in use throughout the world.

This paper will proceed in Section 2 with a description of the challenges for performing entity recognition in subject headings. Section 3 summarizes the state of the art in entity recognition, and Section 4 follows with a description of our approach and details of its implementation. Section 5 presents the evaluation procedure and the obtained results. Section 6 concludes and discusses future work.

2 The Problem

Subject headings present a scenario with particular characteristics for the application of information extraction. To make all the concepts within a subject heading available for machine processing with full semantics, they need to be recognized through named entity recognition techniques.

The available named entity recognition techniques, when applied to subject headings, are unable to reliably identify these entities. These techniques are dependent on the grammatical evidence provided by well formed sentences. In subjects heading such grammatical evidence is not available, since the headings are a concatenation of simple textual references to concepts. The following are some examples of geographic subject headings from the LCSH:

[2] http://id.loc.gov/authorities/
[3] http://www.cs.vu.nl/STITCH/rameau/
[4] http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm

- East Frisian Islands (Germany)
- Potsdamer Platz (Berlin, Germany)
- Danube River--Navigation
- Québec (Québec)--History--French and Indian War, 1755-1763
- United States--History--Civil War, 1861-1865--Propaganda
- Cass Lake (Cass County and Beltrami County, Minn. : Lake)
- Portugal--Economic conditions--20th century
- Portugal--History--Revolution, 1974

In these examples we can observe the heterogeneity of the structure of subject headings. Some delimiting punctuation ("--") is used between the main concepts but they do not provide any clues about the type of the entities that they delimit.

The desired output result, of the entity recognition process, is the location of the entities and the identification of their type. Table 1 illustrates the desired output, as annotated subject headings for entities of the types: geographical entity, topics, time periods, and historical periods.

Table 1. Examples of entity recognition in subject headings

Québec (Québec)--History--French and Indian War, 1755-1763
[GEO Québec] ([GEO Québec])--[TOPIC History]--[HISTORIC French and Indian War], [TIME 1755-1763]
Portugal--History--Revolution, 1974
[GEO Portugal]—[TOPIC History]--[HISTORIC Revolution], [TIME 1974]
Potsdamer Platz (Berlin, Germany)
[GEO Potsdamer Platz] ([GEO Berlin], [GEO Germany])

Several approaches can be adapted for this particular scenario of entity recognition. Similar problems have been addressed in many fields, such as bioinformatics, computational linguistics and speech recognition [3, 5, 6]. Perhaps the most similar has been the citation matching problem [7] where entity recognition is based on structural characteristics of the text instead of grammatical evidence.

In our work we analysed the available techniques and applied a particular one. The chosen technique is better adapted to capture the structural characteristics of the subject headings, and of the entities in them. We also aimed at making our technique language independent, so that our work is generally applicable to any subject heading system.

3 Related Work on Entity Recognition

The Named Entity Recognition task, as proposed by the Natural Language Processing community, refers to locating and classifying atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, etc. [8]. Current solutions can achieve

near-human performance, effectively handling the ambiguous cases (e.g., the same proper name can refer to distinct real world entities), achieving F-score accuracy around 90%. A recent survey of this area is available in [9].

Initial approaches, which are nonetheless still commonly used, were based on manually constructed finite state patterns and/or collections of entity names [9]. In general, the pattern-based approaches attempt to match against a sequence of words in much the same way as a general regular expression matcher. However, named entity recognition is considered as a typical scenario for the application of machine learning algorithms, because of the potential availability of many types of evidence, which form the input variables for the algorithms [5].

A major factor supporting the use of machine learning algorithms for entity recognition reasoning is their capacity to adapt to each case. Thus, they can be deployed with greater flexibility on distinct corpus from different domains, languages, etc. Different types of text analysis methods make available several types of evidence on which to base the named entity recognition reasoning. But not all evidences will be present in every corpus, and not all text analysis techniques will be able to identify the same types of evidence. Therefore, the capacity of machine learning algorithms to adapt to each case make it a very good solution for entity recognition, which is supported by the rising trend in usage of machine leaning in this research area [9].

Two particular types of supervised machine learning algorithms have been successfully used for entity recognition. Early applications applied classification algorithms, which basically classify words, or groups of words, according to their entity type. Some examples are Support Vector Machines [10], Maximum Entropy Models [12] and Decision Trees [11].

Nevertheless, in entity recognition, as in other natural language related tasks, the problem was shown to be better solved with sequence labelling algorithms. The earliest sequential classification techniques applied to entity recognition were Hidden Markov Models [13]. However this technique does not allow the learning algorithm to incorporate into the predictive model the wide range of evidence that is available for entity recognition. This limitation has led to the application of other algorithms such as the Maximum Entropy Markov Model [5] and Conditional Random Fields [4]. Conditional Random Fields is currently the technique that provides the best results for entity recognition. It has sequence classification learning capabilities together with the flexibility to use all the types of evidences that entity recognition systems can gather [14].

4 The Approach

We opted for a sequence labelling approach for recognizing entities in subject headings. The core of our approach lies in identifying the most likely sequence of labels for the words and punctuation marks in any given subject heading. The labels used correspond to the target four entity types (geographical entities, topics, time periods, and historical periods) plus a label for "*not an entity*".

In the remainder of this section we will shortly introduce the predictive model for sequence labelling used in our approach, and then describe the specific features for building our model. In the last subsection we describe the implementation of the overall system.

4.1 The Base Predictive Model

We use as a basis the conditional models of conditional random fields (CRF) which define a conditional probability $p(y|x)$ over label sequences given a particular observation sequence x. These models allow the labelling of an arbitrary sequence x' by choosing the label sequence y' that maximizes the conditional probability $p(y'|x')$.

The conditional nature of these models allows arbitrary characteristics of the sequences to be captured by the model, without requiring previous knowledge, by the modeller, about how these characteristics are related.

A CRF is an undirected graphical model that defines a single log-linear distribution over label sequences given a particular observation sequence. The probability of a particular label sequence y given observation sequence x is a normalized product of potential functions, each of the form [4]:

$$\exp(\sum_j \lambda_j t_j (y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k (y_i, x, i))$$

Where $t_j (y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence; $s_k (y_i, x, i)$ is a state feature function of the label at position i and the observation sequence; and λ_j and μ_k are parameters estimated during supervised training.

When defining feature functions, a set of features $b(x, i)$ is created from the observation sequence. The modeller should choose features which capture those characteristics of the empirical distribution of the training data that should also hold for the model distribution.

Each state feature function uses the value of one of these observation features $b(x, i)$ depending on the current state. Similarly, transition feature functions will use the value of the feature depending on both the previous and current state.

4.2 Features for Subject Headings

For our specific problem of recognizing entities in subject headings, the words and punctuation marks (tokens) of the subject heading form our sequence. Based on the tokens, we defined a set of features that express the major characteristic of the representation of the entities, and would allow the construction of a general predictive model.

We opted to use only features that are language independent, so that the predictive model could be applied to subject headings systems in other languages than the one used for building it. For this reason the words in the subject headings are not used themselves as features, as typically is done in natural language processing. Only characteristics of the words, which we considered relatively language independent, are captured by the features. The following features are used:

$$isWord(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a word} \\ 0 & \text{otherwise} \end{cases}$$

$$isNumber(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a number} \\ 0 & \text{otherwise} \end{cases}$$

$$isCapitalizedWord(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a word and the first letter is capitalized} \\ 0 & \text{otherwise} \end{cases}$$

$$isInitial(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ has just one letter} \\ 0 & \text{otherwise} \end{cases}$$

$$isTinyWord(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a word with character length == 2} \\ 0 & \text{otherwise} \end{cases}$$

$$isSmallWord(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a word with character length >2 and <= 4} \\ 0 & \text{otherwise} \end{cases}$$

$$isYear(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a number that maybe a representation of an year} \\ 0 & \text{otherwise} \end{cases}$$

$headingSection(x, i)$ =Number of previous "--" separators

$$isWhitespace(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is a whitespace} \\ 0 & \text{otherwise} \end{cases}$$

$$isHyphen(x, i) = \begin{cases} 1 & \text{if the token at position } i \text{ is an hyphen} \\ 0 & \text{otherwise} \end{cases}$$

In addition, other features, similar to *isHyphen*(x, *i*), were defined for other punctuation marks: coma, colon, semicolon, period, underscore, open bracket, close bracket, open square bracket, close square bracket, apostrophes and quotation marks.

Additional feature functions are defined in similar way as the previous ones, but they refer to previous or following tokens, instead of the current one. The features: *isWord, isNumber, isWhitespace,*and all *is"PunctuationMark"* are applied also regarding the preceding three tokens, and the following two tokens.

In total, the CRF predictive model is based on 96 features.

4.3 Implementation

The implementation of our approach has eight main components, which are shown in Figure 1. Together they provide the functionality to create an entity recognizer from training data, to evaluate the results of entity recognition, and to process complete subject heading systems represented in SKOS.

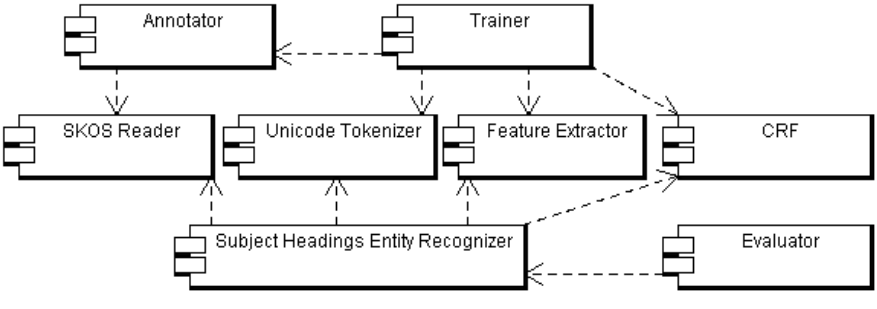


Fig. 1. Component diagram

The Unicode tokenizer component provides the basic text processing to transform the source text into a sequence of tokens (words, punctuation marks, numbers, etc). Although natural language processing tools typically deploy language specific tokenization, we have opted for language independent tokenization since it has also been applied successfully in previous experiments [15]. In particular, we opted for the language independent word breaking rules of Unicode [16] implemented in Java in project ICU - International Components for Unicode[5].

The SKOS Reader component provides the support for reading subject headings from SKOS representations of subject heading systems, allowing their processing by the Entity Recognizer component, and by the Annotator for creation of annotated headings.

The CRF component is provided by the Java implementation in the MALLET - Machine Learning for Language Toolkit [17]. The features for the CRF model, described in Section 4.2, are implemented by the Feature Extractor component.

The Annotator component provides the functionality to read and write annotated subject headings, and to aid a user in creating annotated data sets. It allows a user to correct results from the Entity Recognizer and add them to an annotated data set.

The Trainer component allows the creation of CRF predictive models by supervised training on annotated data sets.

The Evaluator component provides the mechanisms to evaluate the results of the entity recognition process on annotated data sets. It provides functionality for performing cross-validation tests based on the *exact-match* evaluation method for entity recognition described if the following section.

The Subject Heading Entity Recognition component provides an application programming interface for recognizing entities in subject headings. It coordinates all other components necessary for executing the entity recognition and related operations.

The complete system is packaged as a Java jar library. It can be integrated in other applications as a library, or executed from a command line.

5 Evaluation

An evaluation of the recognition technique was performed on a subset of the Library of Congress Subject Headings. A random selection of 800 subject headings was made from subject headings whose main concept is geographic. The subject headings were manually annotated. All entities in the label of the concept were identified and annotated with the corresponding type: geographical entities, topics, time periods, and historical periods. Table 2 summarizes the amount of entities found for each entity type. In total, 1985 entities were found in the 800 subject headings, resulting on an average of 2.48 entities per subject heading.

Table 2. The LCSH evaluation data set

Subject Headings	Geographical Entities	Topics	Time	Historical Periods	Total Entities
800	1348	371	200	66	1985

[5] `http://site.icu-project.org/`

For the evaluation method we have chosen the *exact-match* method, which has been used in several named entity recognition evaluation tasks, such as the Conference on Language Resources and Evaluation [19] and the Conference on Natural Language Learning [20, 21].

In the *exact-match* method, an entity is only considered correctly recognized if it is exactly located as in the manual annotation. Recognition of only part of the name, or with words that are not part of the name, is not considered correct. The following measures are taken:

- Precision: the percentage of correctly identified entities in all entities found;
- Recall: the percentage of entities found compared to all existing entities;
- F_1-measure: the weighted harmonic mean of precision and recall, where recall has the same importance of precision.

The evaluation was performed as a cross-validation test, which involves partitioning the evaluation data set into complementary subsets of the data set, testing the classifier on one subset, while training it on the remaining subset. Ten-fold cross-validation was performed using different partitions, and the validation results were averaged over the ten runs. The results obtained, broken down by entity type, are show in Table 3.

Table 3. Measured precision, recall and F1-measure, using 10-fold cross-validation

Entity Type	Precision	Recall	F_1-measure
Geographical entities	0.981	0.978	0.980
Topics	0.981	0.970	0.976
Time	0.985	0.985	0.985
Historical Periods	0.942	0.985	0.963
All Entities	**0.980**	**0.978**	**0.979**

We consider the results obtained to be good indication that entities can be reliably recognized in subject headings, in a language independent way, and that the CRF based predictive model was able to capture the patterns in the data, achieving an overall F_1-measure of 0.979.

Our analysis from the observation of the cases where the entities were not correctly recognized, indicate that with some simple language specific features (such as using a feature identifying the word "*and*") or with the use of dictionaries of entity names, potentially better results could be achieved. This kind of features can be used in scenarios were language independence is not necessary.

6 Conclusions and Future Work

This paper described a named entity recognition approach for subject headings based on a machine learning algorithm for sequence labelling. The approach was designed to be language independent, having in mind its general applicability to subject heading systems in any language. In our evaluation, our approach achieved an F_1-measure result of 0.979 in recognizing geographical entities, topics, time periods and historical periods.

This work is a first step towards improving the semantic richness of the concepts represented in subject heading systems available in SKOS as open linked data. The positive results of our approach gives us a good foundation towards establishing outgoing links from these open data sets into other widely used open linked data sets, such as Geonames[6] and DBpedia[7]. Therefore, our future work will address the entity resolution problem of the recognized entities into relevant open data sets, which describe the types of entities found in subject headings. As a result links between data sets can be established according to the principles of open linked data.

We also expect that, once the entity resolution problem is addressed, the final outcome can be used to automatically establish links between subject headings systems in different languages, such as the work carried out in the MACS project[8] and related research [22].

Acknowledgments. This work was supported in part by the Europeana Connect project[9], which is co-funded by the European Commission programme eContentplus.

References

Hoerman, H.L., Furniss, K.A.: Turning Practice into Principles: A Comparison of the IFLA Principles Underlying Subject Heading Languages (SHLs) and the Principles Underlying the Library of Congress Subject Headings System. The Haworth Press, Inc., Cataloging & Classification Quarterly 29(1/2), 31–52 (2000)

Miles, A.J., Matthews, B.M., Wilson, M.J.: Core RDF Vocabularies for Thesauri. SWAD-Europe Deliverable 8.1 (2001)

Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)

Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning (2001)

McCallum, A., Freitag, D., Pereira, F.: Maximum entropy Markov models for information extraction and segmentation. In: International Conference on Machine Learning (2000)

Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Signal Processing Series. Prentice-Hall, Inc., Englewood Cliffs (1993)

Wellner, B., McCallum, A., Peng, F., Hay, M.: An Integrated, Conditional Model of Information Extraction and Coreference with Application to Citation Matching. In: UAI 2004 Proceedings of The 20th Conference On Uncertainty In Artificial Intelligence (2004)

Rijsbergen, C.J.: Information Retrieval. Butterworth, London (1979)

Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Linguisticae Investigationes 30 (2007)

Ravin, Y., Wacholder, N.: Extracting Names from Natural-Language Text (1997)

Mikheev, A.: A Knowledge-free Method for Capitalized Word Disambiguation. In: The 37th Annual Meeting of The Association for Computational Linguistics, pp. 159–166 (1999)

Silva, J., Kozareva, Z., Gabriel, J., Lopes, P.: Cluster Analysis and Classification of Named Entities. In: Proceedings Conference on Language Resources and Evaluation (2004)

[6] http://www.geonames.org/

[7] http://dbpedia.org

[8] https://macs.hoppie.nl/pub/

[9] http://www.europeanaconnect.eu/

Bikel, D., Daniel, M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a High-Performance Learning Name-finder. In: Proceedings of the Conference on Applied Natural Language Processing (1997)

Settles, B.: Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In: Proc. Conference on Computational Linguistics, Joint Workshop on Natural Language Processing in Biomedicine and its Applications (2004)

Yamashita, T., Matsumoto, Y.: Language independent morphological analysis. In: Proceedings of the Sixth Conference On Applied Natural Language Processing, pp. 232–238. Association for Computational Linguistics, Seattle (2000)

The Unicode Consortium: Unicode Text Segmentation (2010), http://www.unicode.org/reports/tr29/

McCallum, A.: MALLET: A Machine Learning for Language Toolkit (2002), http://mallet.cs.umass.edu

Lopes, M.I., Beall, J. (eds.): Working Group on Principles Underlying Subject Heading Languages, IFLA Section on Classification and Indexing: Principles Underlying Subject Heading Languages (SHLs). International Federation of Library Associations and Institutions (1999)

Sekine, S., Isahara, H.: IREX: IR and IE Evaluation project in Japanese. In: Proc.Conference on Language Resources and Evaluation (2000)

Sang, T.K., Erik, F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings Conference on Natural Language Learning (2002)

Sang, T.K., Erik, F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings Conference on Natural Language Learning (2003)

Isaac, A., Matthezing, H., Schlobach, S., Zinn, C.: Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies. Library Review 57 (2008)

Towards Cross-Organizational Interoperability: The LIDO XML Schema as a National Level Integration Tool for the National Digital Library of Finland

Riitta Autere[1] and Mikael Vakkari[2]

[1] Finnish National Gallery, Kaivokatu 2, FIN-00100 Helsinki, Finland
[2] The National Board of Antiquities, PO BOX 913, FIN-00101 Helsinki, Finland
riitta.autere@fng.fi, mikael.vakkari@nba.fi

Abstract. The Finnish National Digital Library (NDL) project aims to improve online accessibility and usability of digital content held by libraries, museums and archives. The lack of standardized metadata and numerous different collection management systems without sufficient set of technical standards in the museum sector led us to create a set of instructions and a template mapping of the Lightweight Information Describing Objects (LIDO) XML schema. This national LIDO schema for museum sector described in our paper is unique both in coverage in museum object types as well as number of institutions using it. A common schema presents heterogeneous metadata uniformly, thus enabling easy retrieval, browsing and versatile linking between different object types as well as data fields. In the pilot phase we have mapped the three most commonly used Finnish collection management systems with three different metadata formats to the top level LIDO schema.

Keywords: interoperability, LIDO, metadata standards, XML schemas.

1 Introduction

The National Digital Library (NDL) is a project launched by the Ministry of Education and Culture of Finland in 2007 [1]. The aim of the project is to improve the availability and usability of the digital content held by libraries, archives and museums as well as to develop a long-term preservation solution for the digital material. The NDL is part of the development of national electronic services and infrastructures in Finland and one of the public administration projects defined in the *Ubiquitous Information Society Action Programme implementing the Government Resolution on the Objectives of the National Information Society Policy 2007-2011* [2].

The NDL User Interface Project goal is to provide a homogenous view to diverse digital records and make different elements of this view to operate for the users benefit by using the elements of the descriptive metadata harvested from cultural organizations' collection management systems and creating links and functionalities between different data elements. From the end-user's point of view the user interface provides tools for easy and seamless movement from one digital resource to another

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 62–68, 2011.

while allowing users to expand the search results to a wider information network. These aspects offer us a strong cross-organizational perspective on development of tools for the digital content delivery.

This paper formalizes our work to create a standard tool and a set of instructions for cultural heritage organizations to use in digital content delivery, some motivations behind our work, and the first results with some wider prospects to continue with.

With content we mean content as defined in *The Digital Library Reference Model* version 1.0: Content includes all the data and information, all forms of information objects handled and made available, composed of sets of information objects organized in collections [3].

The IEEE Glossary defines interoperability as the ability of two or more systems or components to exchange information and to use the information that has been exchanged [4]. Interoperability is observed in this paper from the view of content, functionality and architecture. We agree with Athenasopoulous et al.'s view "interoperability being crucial to improving efficiency and effectiveness" [5].

With content interoperability we mean the system's ability to directly share content and metadata from different organizations and the ability to support, and create whether possible, the connections and links between different data elements of this data. We are aware that there are several generic concerns regarding crossorganizational descriptive metadata interoperability.

2 The Existing Data Structure in the Finnish Museums

The most problematic issues at the starting point of our work were related to the amount of different collection management systems in the Finnish museums and the data formats and structures used in them. The museums have traditionally organized themselves in small groups or consortia and developed collection management systems and metadata formats of their own to support the handling of their digital resources. Depending on the definition of a museum collection management system there were more than 30 different collection management systems used in 157 professionally held museums in Finland in 2009 [6,7,8]. Only one of these systems is based on *CIDOC Conceptual Reference Model,* while some are adaptations from different standards such as SPECTRUM, CDWA and museumdat, but most of them are not based to any standard data format [9].

To export digital content from this vast number of heterogeneous collection management systems to a single digital library solution seemed to be a major challenge during the NDL implementation project. The import pipes of the digital library system and boundaries created by the various data structures and formats as well as the different processes of data management and delivery manifested as an unmanageable scenario of dozens of variations of the pipes and normalization rules at the early days of the implementation project.

3 Towards Cross-Organizational Interoperability

The challenge in creating a real cross-organizational interoperability between different organizations' data structures and digital content includes several gaps which need to be

bridged. One major gap lies between the different data formats used by museums and the delivery of the data to library-oriented software. Another gap lies between the data exchange formats used in museums and those in libraries, e.g. MARC, FINMARC and Dublin Core.

Bridging these gaps is the one of the main tasks of our work. Our aim is to make these gaps smaller and the export/import process and normalization smoother and easier by developing a unified data structure and a data exchange format suitable for delivering digital content from heterogeneous museum collection management systems to a single system. The data delivery process will be ensured with establishing tighter connections with different organizations' digital content with a suitable set of instructions and guidance for the data exchange format.

With a unified data structure and format for digital content delivery from museum collection systems, a set of instructions to use this structure and format and the support and guidance from the two national museum sector institutions, we are aiming towards high integration between various data elements of the digital content held by Finnish museums.

An organized, harmonized and standardized data exchange format for museum information would be an asset for museums and the delivery of the information in unified format would benefit not only the museums but also the NDL project itself.

The process towards cross-organizational interoperability began in January 2010 by tests with Dublin Core as a data exchange format for both museums and libraries. These tests held during the NDL system implementation pilot project immediately showed after a single test session that Dublin Core lacked the depth needed for the metadata delivery from museums. A more versatile format was needed.

4 The National LIDO XML Schema Template

Our aim was to define an XML schema for the exchange of museum collection information and digital content to digital library and web portal solutions. The problems perceived during the implementation of the NDL project harvesting and normalization processes needed to be solved before these processes were launched as a large scale, nationwide operation for the cultural heritage sector.

4.1 The Lightweight Information Describing Objects (LIDO)

The LIDO (Lightweight Information Describing Objects) XML schema [10] is created within a specific working group of *CDWA Lite Advisory Committee* and the *Documentation Committee of the German Museum Association* as a schema that meets the requirements articulated by formats of CDWA Lite and museumdat, as well as the feedback received from the community of information and technology professionals [11]. Early versions of the schema and its declarations were provided for professional feed back during 2009. This schema seemed a suitable candidate for resolving our problematic issue for data exchange in Finnish museums after the tests run with Dublin Core.

LIDO version 0.9's documentation was researched by us under the late spring and early summer 2010. The data structure seamed rich and scalable enough, wrappers and

sets logical and suitable for the descriptive metadata of both cultural historical objects and art works. Declaration was clear, easy to understand and versatile. After some consideration we ended up to testing LIDO as the data exchange format for delivering the digital content of the Finnish National Gallery's collection information and soon after that the National Board of Antiquities' collection information. Both systems' data was mapped to the template and the results were encouraging. In October 2010 we had the first large scale cross-organizational mapping template finished and available for other organizations. Just for few weeks before a new LIDO version, LIDO 1.0, was launched in *ICOM 2010 General Conference* in Shanghai, China. The LIDO 1.0 had some modifications and small structural differences to the previous version, thus we re-mapped our template to it.

4.2 From a Model Schema to the National Level Template

Delivering metadata resources beyond the boundaries of a museum collection management system requires that the receiving system is both able to interpret and process the data elements and normalize and index the information as the delivering organization intends.

To make the data management easier during the NDL delivery and end user interface implementation project, the museum organizations involved agreed to deliver their data in the LIDO format. This decision cut down the estimated amount of different mapping rules ('pipes'), but left a set of normalization rules of different unique adaptations of the general LIDO schema. After a policy decision on the use of one common schema, several museums with three different collection management systems and three different mappings of the LIDO schema, ended up using one shared LIDO XML schema with a single set of normalization rules. The systems still have individual OAI-PMH API's and pipes for the harvesting but the overall mapping is done to the shared schema thus eliminating the need for multiple individual mappings.

With a detailed study of the content and the use of each element in the different museum collection management systems we defined all equivalent elements present in the digital content available for harvesting. Identification of the content of each describing metadata element and defining the relationship between these elements with equals in different collection management systems was a base for a national level data exchange format. This identification process was also a step towards crossorganizational interoperability in the context of the NDL.

The harvested data is normalized and indexed into the digital library system and published in the End User Interface for fast retrieval and clear presentation. The normalization rules need only be applied once according to the national level LIDO schema making multiple mappings unnecessary and thus saving a great amount of time and effort.

The nationwide LIDO schema template and the local mappings to it have been under development during the NDL pilot phase, and have been re-defined several times during the project. The schema is aligned with the content configuration parameters of the NDL distribution and user interface system and it includes all the data elements describing museum object and its connections to other digital resources. Due to the functionality of the digital library system chosen for the NDL the elements with ontological enrichments or annotated information has not been included to the top level schema yet.

Administrative metadata of the LIDO schema includes technical characteristics of the digital resources, source metadata, such as the information about the source where the resource has been produced, rights management metadata with all the details of copyright information, use restrictions and license options. Unfortunately the importance of the rights metadata is considerably often neglected or not understood as a vital element of the administrative metadata of a digital resource.

After having specified the necessary elements and attributes from the rich original LIDO documentation, we still have to meet the challenge of publishing these specifications in a way that can be understood and implemented among the diverse museum collection management systems in Finland. This task requires shared mechanisms, processes and defined best practices.

The challenge of mapping the original LIDO schema to the variety of existing data formats and models is closely related to the reality of the collection management systems' existing non-standardized conventions and structures in Finland.

The Finnish museums now have a common XML template for data delivery in LIDO 1.0 format. The template will provide tools for harmonizing and delivering standardized data from heterogeneous collections stored in various types of collection management systems by using a national unified LIDO XML schema for the whole Finnish museum sector.

4.3 Flexibility as a Benefit and a Drawback

The scalability of the LIDO schema is excellent. The absence and the placement of some element types, especially events and procedures of conservation, and the incoherence between the elements of the LIDO schema and the existing cataloging methods and fields in the museum collection management systems in Finland contributed to some compromises.

The most frequent problem type occurred with LIDO import/export processes during the NDL implementation project was caused by the LIDO schema being even too flexible and scalable in relation to our current systems. The systems or portals harvesting or importing digital content in LIDO format are seldom capable of utilizing the whole benefit of the detailed information delivered to them. Neither the NDL management system nor the ESE format developed for digital content delivery into the *Europeana* portal [12] could use the richness of the LIDO format.

At the moment the LIDO schema template contains much more elements andattributes than needed with data exports for NDL's use. The template is not constructed and tailored only for the needs of the NDL, but to meet data exchange needs in various future projects as well.

5 Cross-Organizational Interoperability of the LIDO Template and the National Digital Library of Finland's End User Interface

By implementing a national level LIDO schema as the only harvesting format, the data gathered from various collection management systems in several differing formats is easy to normalize with one set of normalization rules in the digital library management system for the use in the end user interface.

There is no need for the data provider, i.e. the organization delivering the digital content, to familiarize himself with several different methods and interfaces for distributing data when the data elements and information included in the metadata from different sources is uniform, understandable and easily accessible.

The descriptive metadata delivered by museums can also be linked with the data from libraries and archives. For example, a museum object which has a related publication, say a research article or an exhibition catalogue, and a data element in the describing metadata containing the name and/or the ISBN of that publication can be easily be linked to the publication located within the library data. These functionalities allow organizations taking part in the NDL to provide a new kind of information set to their customers in a new context by presenting both the object of interest and other relevant information related to it together instead of just showing a set of objects as a result of the query. There is also a possibility to enrich the descriptive metadata of the object with direct links to other related information or published sources, for example from the creator (e.g. architect, artist or author) name field to the on-line accessible biographical information or article.

6 Conclusions

A creation of the national level XML schema template initiated from an urgent and unforeseen need for a suitable data exchange format. The benefits recognized at the early stage of the project were the fertile co-operation between museum institutions and the saving of both time and resources by doing things together with simple but well-defined set of standardized tools.

The LIDO schema created just months before offered a new possibility to resolve the lack of standardized data format for the digital content delivery from the Finnish museum collection systems.

The common XML template with its crosswalk tables will provide tools for harmonizing and delivering standardized data from heterogeneous collections stored in various types of collection management systems by using a national unified LIDO XML template for the whole Finnish museum sector. The template is already in use with a standard OAI-PMH API implemented for the museum systems during the NDL project. We have also enhanced the reusability of the data and saved resources by adapting this XML template for other data delivery needs to external systems, such as Europeana.

A key part of the project relates to the implementation of widely accepted and used standards and best practices to Finnish museum collection management systems and processes. The use of the national level template requires good will and an adoption of a shared set of technical standards and guidelines by all the museums delivering their digital content. We can't make the use of the template mandatory, only recommend and advice the organizations to use it.

The expected benefits of this template are possibilities for additional features supporting data usage and exchange, for example to take advantage of the geospatial information included with the objects, ontology utilization, and the museum collection management system developers being able to identify and create appropriate implementations to their products. Also more advanced approaches to the

data interoperability among cultural heritage organizations and on-demand integration can be better supported.

The presented template is still undergoing piloting together with the NDL system and has so far been tested with three different collection management systems. In next phase we expect there to be more collection management systems adapting the template and some new distribution portals importing the content delivered in this format.

In this paper we have described one of the elements of the interoperability issues of the National Digital Library of Finland. After providing our first version of the national level XML schema template, we envision a broader use of the schema in Finland and the original LIDO schema building an easy-to-exchange data format for cultural heritage institutions. So far this project has started several fruitful discussions on the state of the museum collection management systems and standards used in Finland. We are looking forward to hear users' and system developers' comments, and hope this project can be a contribution to the next version of the LIDO schema.

References

1. The National Digital Library, http://www.kdk.fi/en/
2. Ubiquitous Information Society Action Programme implementing the Government Resolution on the Objectives of the National Information Society Policy 2007-2011. Yliopistopaino (2008), http://www.arjentietoyhteiskunta.fi/files/73/Esite_englanniksi.pdf
3. Athanasopoulos, G., Candela, L., Castelli, D., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G., Ross, S.: The Digital Library Reference Model Ver 1.0, DL.org: Coordination Action on Digital Library Interoperability, Best Practices and Modelling Foundations - Project Number: 231551, 20 (2010)
4. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. Institute of Electrical and Electronics Engineers, New York (1990)
5. Athanasopoulos, G., El Raheb, K., Fox, E., Kakaletris, G., Manola, N., Meghini, C., Rauber, A., Soergel, D.: A Framework for Digital Library Function Description, Publication, and Discovery: A prerequisite for interoperable digital libraries. In: Workshop on Making Digital Libraries Interoperable: Challenges & Approaches, vol. 20. Glasgow, Scotland (September 2010)
6. Hongisto, V.: Katsaus museoiden kokoelmanhallintajärjestelmiin. Museovirasto, Helsinki, http://www.museoliitto.fi/doc/Hongisto100209.pdf
7. Ekosaari, M.: Tietokoneet museotyössä – Suomalaisten taidemuseoiden tiedonhallinnan historia ja nykytila. Tampere (2009)
8. Koivisto, T., Kaukonen, M. (eds.): Museotilasto Finnish Museum Statistics, vol. 2. Museovirasto, Helsinki (2009)
9. Museoiden kokoelmanhallintajärjestelmät. Nykytila ja vaihtoehtoiset kehitysmallit. Selvitystyön loppuraportti. CSC – Tieteen tietotekniikan keskus Oy, Espoo
10. Coburn, E., Light, R., McKenna, G., Stein, R., Vitzhum, A.: LIDO – Lightweight Information Describing Objects, Version 1.0. ICOM International Committee of Museums (2010)
11. CIDOC Data Harvesting and Interchange Working Group, http://www.lido-schema.org
12. Europeana, http://www.europeana.eu

Supporting FRBRization of Web Product Descriptions

Naimdjon Takhirov, Fabien Duchateau*, and Trond Aalberg

Norwegian University of Science and Technology NO-7491 Trondheim, Norway
{takhirov,fabiend,trondaal}@idi.ntnu.no

Abstract. The FRBR model has the potential for new services and discovery techniques for cultural items such as books, movies and music. In this paper, we present an approach to interpret descriptions found in Web resources and identify the FRBR entities these pertain to. To verify the resulting set of FRBR entities, we have used the Linked Open Data and the verifications have been validated by a group of experts. The results of this work demonstrates applicability of FRBR in a new context and establishes a firm basis for further exploitation.

1 Introduction

The entity relationship model proposed in the IFLA Functional Requirements for Bibliographic Records (FRBR) [6] provides a framework for the entities and relationships that are of interest to end users of metadata. The model builds upon current practice and understanding of what commonly is described in metadata, and defines a formal framework for explicit statements about the entities and relationships that such descriptions pertain to. The interpretation or conversion of bibliographic records is a topic explored in different projects [4,5,2,7]. Projects so far have mainly focused on library catalogs, but the FRBR model is generally accepted as sufficiently generic to serve as a conceptual framework for a broad range of metadata related to cultural items. The major benefit of the FRBR is that it can be used to describe intellectual and artistic contributions at different levels of abstraction. The model enables collocation of entities based on their intellectual equivalence and the relationships between the entities provides a network-based structuring of the entities described in metadata.

For many users the Web is the primary source of information and the total amount of data available online is far larger than the one stored in library catalogs. However, this Web data is often neither well structured nor machine-interpretable, although the emergence of the Semantic Web aims at tackling this issue. For a large portion of Web data there is, unfortunately, no easy transition to the Semantic Web. Simply transforming from one format to another, which is a syntactic approach, does not automatically enable semantic interoperability and input data often needs to be reinterpreted into entities and properties that make sense as well as transformed.

* The author carried out this work during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 69–76, 2011.

In this paper, we propose to fill the gap between these two worlds, using the FRBR model for product information found on the Web. The Web contains a substantial amount of resources that describe products of creative or artistic endeavor such as online stores and community sites. We present an approach to interpret such information and identify the entities of the FRBR model. We advocate that such a representation would enable websites (e.g. e-commerce) to better organize and exploit those products.

2 Functional Requirements for Bibliographic Records

FRBR is a conceptual model of the bibliographic universe published around a decade ago [6]. Intellectual and artistic contributions are modeled in multiple levels of abstraction using the entities: **work**, **expression**, **manifestation**, and **item**. Figure 1 depicts this hierarchy. The three-part epic by J.R.R. Tolkien "The Lord of the Rings" is an abstract work encompassing "The Fellowship of the Ring", "The Two Towers", and "The Return of the King". Work represents a distinct, intellectual or artistic creation. Each of these three works has been translated and published in number of languages and each of those translations/editions is an expression in the FRBR model. This is illustrated by the realization of "The Two towers" in two different languages: the original English version "The Two Towers" and a Norwegian translation "To Tårn". The paperback format in original language (English) published by Mariner Books in 2005 is regarded as a manifestation in FRBR terms. In our example, the paperback edition published in September 2003 and June 2005 are thus regarded as the two separate manifestations of the same (English) expression. Finally, the physical book that one can hold in his hand is an item.

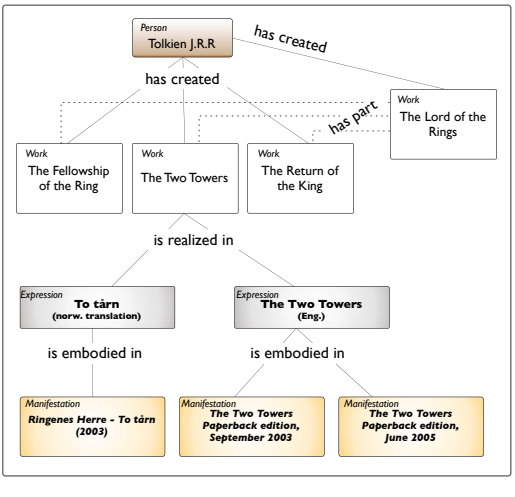


Fig. 1. A Fragment of Lord of the Rings FRBR Work by J.R.R. Tolkien

A person (or corporate body), sometimes also referred to as *actor*, in the FRBR model, is an individual responsible for the creation or realization of a work (e.g., as an author, an illustrator, a translator, etc.). At the same time, this type of entity can be the subject of a work.

Additionally, the FRBR model provides **a set of relationships** between entities beyond the basic relationships shown in Figure 1. This feature helps to cluster a work and its related entities (e.g., an adaptation of a book in a movie), ultimately leading to better user experience when searching and exploring a collection.

3 FRBRizing Web Product Descriptions

One of the main difference between existing FRBRization approaches and our work deals with the input data. Our FRBRization process takes as an input descriptions of products found on the Web, specifically products sold by e-commerce websites (e.g., Amazon). Information about products tend to have different properties from their bibliographic record counterparts found in library catalogs. A first difference is that products do not have the same structural pattern as MARC records, and they are stored in a variety of formats. A second one deals with the identification of products, which are unambiguously referenced by URI, thus providing a basis for reuse and exchange [3]. Additionally, e-commerce websites usually provide faceted navigation where the ranked list of results can be filtered on several dimensions. Yet, Web products can be related to FRBR manifestation level, similarly to library catalogs. For example, the 2005 paperback version of the book "The Two Towers" sold for $10.95 at `hmhbooks.com` is a product which is an original English expression of the work "The Two Towers" by Tolkien.

Subsequently, our approach consists of a set of interrelated operations which result in identification of the different FRBR entities. The FRBRized entities are then connected by establishing appropriate FRBR relationships. The FRBRization workflow is illustrated in Figure 2. From the input product descriptions, we first identify the corresponding works (Section 3.1), then we generate related manifestations, expressions (Section 3.2) and actors (Section 3.3). The process of creating relationships between the FRBR entities, presented in Section 3.4, produces the FRBR collection.

3.1 Identifying a Work

The FRBR work is an abstract distinct intellectual or artistic creation. This entity is the cornerstone of the FRBR model and any FRBRization process needs to include a method for identifying the work entities. Description of a single product on the Web reflect the manifestation level, but attributes of expression and work can often be found in the description of the product. For example, the language of the book is an attribute of the expression while the title and author(s) may refer to the original work.

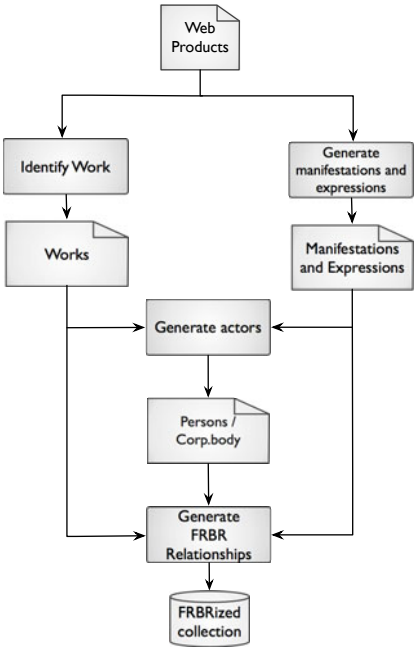


Fig. 2. The FRBRization Workflow

The following techniques can be used to identify a work:

- Creating a work based on title/author and other attributes of the resource if the work has not been created yet; the database of works is then incrementally updated;
- Using an external service to identify a work, e.g. OCLC Classify API[1] for books using ISBN number, ISMN/ISRC music database for music or IMDB API for movies;
- Use z39.50 (or SRW/SRU) protocol to search and fetch the relevant MARC record from publicly available catalogs and then use similar technique to Work-Set algorithm by OCLC [5].

The two latter methods take an identifier as input. On the contrary, the first method requires attributes such as title/author, thus leading to string matching problem. For instance, if the input product description is a translation of a work, we have to make sure that the correct work is discovered.

3.2 Generating Related Manifestations and Expressions

The FRBR expression is often is perceived as the most difficult entity because information that is needed to distinguish between such entities often is

[1] http://classify.oclc.org

ambiguous or missing. On the other hand, because the expression is an intermediary node between work and expression, we are able to generate expressions by clustering the manifestations that are related to a work using any expression level attribute. The main challenge is to discover the related manifestations, and we have identified the following methods to achieve this goal:

- Search for author in z39.50 enabled repositories;
- Use external service (e.g. xISBN[2] or ThingISBN[3], Spotify[4]).

These methods rely on external sources. Note that using the first method implies to employ FRBRization techniques already proposed for MARC records [7,5,1].

From the set of related manifestations, we automatically generate expressions by analyzing attributes pertaining to the expression level. For example, attributes such as language and translator are used to identify expressions. Note that the identifier for an expression is automatically generated at this stage.

3.3 Generating Actors

An actor is a person or corporate body (organization) responsible for the creation or realization of a work. Products available from e-commerce websites usually have information about the responsible for the work, such as author of a book, composer of a music, director of a movie. Generating an actor can be performed using the following methods:

- Create a local authority file or use existing authority files from external sources;
- Search Virtual International Authority File (VIAF) and link actors to VIAF.

Contrary to the first method which is a time-consuming and complex task, the second approach includes a Web-based API and the VIAF collection contains data from many national libraries around the world. At the end of this step, we have generated all the FRBR entities required in our FRBRization process.

3.4 Generating FRBR Relationships

The final phase of the workflow is to generate the actual FRBR entities and establish relationships. Since we have information about each entity from previous steps, this step creates a collection of entities in a specific output format such as a series of SQL statements that can be used to insert into a relational database, HTML, XML, RDF or a simple text file. This step requires that an entity has a unique identifier and can be unambiguously referenced. For manifestations, we already have identifiers. Work and expression entities can be assigned locally generated identifiers since there is no publicly available global unique identifiers for these entities.

[2] http://labs.oclc.org/xisbn/
[3] http://www.librarything.com/api
[4] http://developer.spotify.com/en/libspotify/

4 Experiments

In this section, we demonstrate the use of our approach and the level of quality we have achieved. We have chosen to create our dataset based on search results from Amazon since its database potentially contains a great number of items[5]. Another reason to use Amazon is that some of the sites already implement a feature that is comparable to FRBR model by presenting users with a list of alternative formats. This is, however, solely based on metadata equivalence and is limited to publications that appear under the same title and author.

4.1 Experimental Protocol

Using Amazon's Product Advertising API[6], we have searched for works by the 80 best selling fiction authors extracted from Wikipedia[7]. Due to the constraints set forth by Amazon on the number of requests that can be sent in one hour, we have a limited number of items to the first page of the results set (10 items per page). We have performed an automated search using the *ItemSearch* operation on *Books* index. We excluded items representing kindle edition. We also filtered out the products not solely offered by Amazon ("MerchantId"=Amazon). Additionally, we performed search on Amazon's *Video* and *Music* indexes using previously submitted queries on *Books* index. The attributes made available within these products, among others, are, title, author (director for movies), contributor, ISBN, language, release date. As can be seen from Table 1 (column "# of Input Products"), half of the content of the initial set of products were books. Most of these books are published in English language, but the input products include other languages such as Japanese, Chinese or Russian.

In Section 3.1 we have proposed three methods to identify the work corresponding for a product. To avoid implementing z39.50 protocol and reduce latency, we used the Classify API by OCLC. Classify API is a web service from the OCLC Office of Research that can be used to retrieve information, such as work level title, that is common the group of publications that belongs to the same work (identified by the use of OCLC's workset algorithm). The next step in the workflow is to generate related manifestations and expressions. Since we chose OCLC Classify API to identify work, we had a greater chance of match in the same database. Therefore, to obtain a list of related manifestations, we again used an OCLC Service - xISBN. The xISBN Web service returns ISBNs and other information associated with an individual intellectual work that is represented in the WorldCat catalog.

The next step involves the identification of actors. We chose to link actors (persons and corporate bodies) to Virtual International Authority File (VIAF). VIAF is a joint project of national libraries of several countries and it is hosted by OCLC. VIAF's long-term goal is to include authoritative names from many libraries into a global service that is available via the Web for free. Using VIAF's

[5] A blank search on "Books" generates 32,058,092 items (January 2011).
[6] `http://j.mp/amznAPI`
[7] `http://j.mp/fiction_authors`, as of November 2010

public API, we submitted queries for each contributor in the dataset. We used an average of Monge Elkan, Jaro Winkler and Levenshtein to calculate the similarity in the top 30 hits. The final phase of the workflow was to generate the relationship between the FRBR entities. This is achieved using the identifiers created for the FRBR entities in the previous steps. The final output is a set of RDF files for each entity type.

4.2 Quantitative Results

Table 1 summarizes the results of this experiment. The second column provides the number of Amazon products grouped by product type and the number of actors extracted from these products. In the third column, we show the number of discovered entities during the FRBRization process using Classify API, xISBN and VIAF services. Contrary to what could have been expected, the number of discovered works (739) is less than the number of input products (1216). This occurs because the set of input products contains different products that correspond to the same work (e.g. Norwegian "To tårn" and English "The Two Towers" both corresponding to "The Two Towers" work). We notice that the number of manifestations strongly increased (from 1656 to 28245) because we fetched all manifestations of works provided by the xISBN service. More specifically, we successfully discovered more books and videos while related music and DVDs were more difficult to fetch. The total number of actors we extracted was 2221 while 70% of them were found in VIAF (1569).

The last column describes the number of entities in our FRBRized collection, i.e., after removing unidentified works and actors. The initial set of input products we populated from Amazon contained 1656 items while the number of FRBRized manifestations is **28245**. The FRBRized collection includes works, translations, and movie versions of those works. Out of total 739 generated works, we have obtained a match for **684** works in Classify. The unidentified 55 works were mainly not in English language. Dealing with the actors, the final collection contains **2221** actors since the generated actors based on VIAF were automatically assigned locally generated identifiers. Finally, the following issues were encountered during the experiment:

Table 1. Results of the FRBRization

FRBR Entity	**# Input Resources**	**# Discov. Entities**	**# FRBRized Entities**
Work		739	684
Expression			5074
Manifestation	1656	28245	28245
-Book	856	27588	27588
-Video	102	542	542
-DVD	190	113	113
-Music	508	2	2
Actor	2221	1569	2221

- Search results from Amazon often needed to be cleaned. Some data was deemed as dirty, e.g. if the title referred to a movie rather than a novel such as "The Lord of the Rings: The Return of the King (Widescreen Edition)". Since we performed search on Amazon database, we could not always limit our list to only works by our initial set of authors. This happens because authors could be mentioned in descriptive text of the resource;
- We could not FRBRize the whole set of product descriptions because the Classify service did not have an entry for all requested products. To solve this issue, we could aggregate the results from similar services (e.g. z39.50);
- VIAF had several identical entries for number of authors (e.g., "Arthur Rankin Jr."). In this case, the system chooses higher ranked item and if the score is identical, the item is chosen in random manner.

5 Conclusion

In this paper, we have demonstrated that the use of the FRBR model as a semantic data model is not only limited to library catalogs, but can be applied to product information found on the Web too. The main benefit of FRBRizing product information on the Web is that FRBR provides support for knowledge-like representation of the data enabling a broad support for exploratory interfaces where users are presented with a list of works for each author and can navigate relationships to learn about and find other versions or preferred editions of a given work. In the future, we plan to study support for more complex tasks such as automatic detection and extraction of aggregate works. Our FRBRization process can be further improved by using text analysis techniques. This means we automate the process of identifying entities and establishing relationships. At the application level, we plan to infer interesting relationships between the works linked to LOD.

References

1. Aalberg, T.: A Process and Tool for the Conversion of MARC Records to a Normalized FRBR Implementation. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 283–292. Springer, Heidelberg (2006)
2. Freire, N., Borbinha, J.L., 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., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 267–276. Springer, Heidelberg (2007)
3. Gerber, A., Hunter, J.: A compound object authoring and publishing tool for literary scholars based on the IFLA-FRBR model. International Journal of Digital Curation 4(2) (2009)
4. Hegna, K., Murtomaa, E.: Data mining to find: FRBR? In: 68th IFLA General Conference and Council, Glasgow, Scotland (2002)
5. Hickey, T.B., O'Neill, E.T., Toves, J.: Experiments with the ifla functional requirements for bibliographic records (FRBR). D-Lib Magazine 8(9) (2002)
6. IFLA Study Group on the FRBR. Functional requirements for bibliographic records, final report. UBCIM Publications; New Series 19(1) (1998)
7. Manguinhas, H.M., Freire, N.M.A., Borbinha, J.L.B.: FRBRization of MARC records in multiple catalogs. In: Proc. of JCDL, Gold Coast, Australia (2010)

Assessing Use Intention and Usability of Mobile Devices in a Hybrid Environment

Spyros Veronikis[1], Giannis Tsakonas[1,2], and Christos Papatheodorou[1,3]

[1] Dept. of Archives and Library Sciences,
Ionian University, Corfu, Greece
[2] Library and Information Center - University of Patras
[3] Digital Curation Unit, IMIS, "Athena" Research Center,
Athens, Greece
{spver,gtsak,papatheodor}@ionio.gr

Abstract. During the last decades many information providers, such as libraries, have been collecting, organizing and delivering information in both print and digital format, forming a hybrid information environment. However, exploration of a hybrid information environment does not result in a unified seeking experience, which exploits most effectively the available resources. This paper aims to identify the main factors that influence the adoption of wireless, mobile devices (e.g., smartphones) as a means of integrating the information seeking process in hybrid environments. Therefore it presents a prototype system and an evaluation study that provides an insight about the services design.

Keywords: Digital library evaluation, interaction, mobile devices.

1 Introduction

In a hybrid information environment users can easily search for both printed and digital documents, but usually not simultaneously. For example, when library users are interested in printed material they need to browse among the stacks of books and journals, thus losing the access to the available on-line tools because computer rooms and stacks are not usually found in the same place. Therefore, they are forced to seeking strategies, which raise inconvenience and obscure seamless interaction with the available information spaces. In general users must be able to explore the links among several data items of different format and to do so they need to interact with the two available information spaces, explore them and retrieve the relevant documents. Hence, the information seekers need to actively alter their seeking strategy and parameters whenever they feel that such an action will lead to a better result. In addition, they need to maintain access to the on-line tools, without any restrictions regarding their location. The recent advancements in mobile computing technology have equipped mobile devices such as Personal Digital Assistants (PDAs) and smartphones with increased processing power and memory, high-resolution screens and Internet connectivity. These lightweight devices can be easily transferred and used

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 77–88, 2011.

from anyplace within a hybrid information space, thus allowing users to explore the physical information space while maintaining access to the available digital resources (tools, collections, etc.)

This paper aims at investigating the potential of mobile computing devices as a means of unifying the information seeking processes in both physical and digital spaces. The main interest is oriented to the factors affecting the adoption of mobile technology for the achievement of a seamless information access experience. The identification of the factors and the assessment of the effect size can set the priorities for the improvement of services that are based on mobile devices. The next section describes the related work on the use of mobile devices to enhance exploration of hybrid information spaces and a model used to assess the adoption of computing technology within the context of information services. Section 3 describes the research framework and methodology, while section 4 provides a brief description of a mobile-based prototype system and supported functionalities. Section 5 describes the experimental setup and discusses the corresponding results, while section 6 presents the derived conclusions.

2 Related Work

The potential introduced by mobile computing devices in on-line information services was quickly acknowledged by researchers as a facilitator of (a) submitting a query to a catalog from anyplace and subsequently receiving metadata or full-text information, (b) exchanging short messages via email or RSS applications, (c) creating notes and accessing tables, indexes, graphs and other frequently used material [2]. Satpathy and Mathew [10] report a context-sensitive application that enhances location-aware information retrieval and provide services, which improve the operation of the physical organization. Jones, Rieger, Treadwell and Gay [5] stated that the participants in their study were in favor to communication functionalities between library visitors and information experts. Library visitors expressed their appreciation to be able to receive remote assistance during their information seeking session without needing to refer to physically defined points. Mobile scanners of Radio Frequency Identification (RFID) tags have also caught the attention of developers and researchers for mobile library services, due to easiness and speed in retrieving information from/for tagged objects. Recently Buchanan [4] introduced the concept of the "fused library", where the digital library supports user's information seeking process, triggered by contextual stimuli, as in the case of RFID tags. Buchanan and Pearson [3] presented the architecture of the EmLi prototype that retrieved digital library objects, invoked by RFID tags attached on objects that reside on the physical environment of the users. The authors presented preliminary results from a user study, where it was found that this kind of interaction was positively accepted by the users. Similarly, the users in Lin et al. [7] stated their satisfaction with the services provided in the context of a museum and expressed their interest in extending these towards new collaborative features.

Within the last five years, mobile computing studies in the field of information science have been attracting researchers from the fields of engineering, humanities and psychology. However the previous projects do not tackle the issue of providing a seamless interaction experience to users that seek information in hybrid spaces. In general, navigation, search and retrieval in a hybrid information environment raise many contextual concerns. Within a hybrid information environment the context might include places, users (e.g., modeled by skills, education, needs or preferences), objects (e.g., printed and digital/digitized documents), and activities (e.g., tasks, goals, events). Ryan and Gonsalves [9] further expanded the concept of context to include the mobile device itself. As a result their model includes three context components; participants, hybrid information space and computing device as well as the interactions between them: (a) *Users-information space* which describes the information seeking behavior. The most commonly used information discovery techniques are searching and browsing and due to mobility the users can explore the conventional information space while maintaining access to the available digital resources. Therefore, this interchangeable interaction supports a non-linear information seeking behavior. (b) *Users-device* which refers to the user interface and the tools that support information discovery and is strongly affected by their usability. Specifically, the usability of mobile interfaces is affected by the input modes used (e.g. virtual keyboard or handwriting recognition), the size of the screen and the runtime environment, i.e., processing power, memory, storage capacity, etc. (c) *Device-information space* which describes the transactions and data exchange between the device and components of the hybrid information environment. Proper sensors are used to detect nearby available services or items, which might be transparent to the users. In addition, location-based information can be pushed to the device as the users approach a certain location. Also, optical and proximity sensors can be used to quickly retrieve data from objects, such as books tagged with a barcode or an RFID tag. Although sensors, like RFID tags or QR codes, are subject to proximity measures, they are the only means to bridge the components of the hybrid environment. Mäkelä et al. [8] investigated the factors affecting the mobile interactions with RFID tagged and QR coded objects. They reported significant problems, which were associated with sensors' visual cues, as well as with the mental models that the users have about the potential uses of the sensors.

3 Research Framework

The work presented in this paper seeks in gaining better insight and understanding of the factors known to affect the adoption of technology while exploring the hybrid information space with a mobile device. The research questions formed in this context are:

– can mobile devices be used to bridge the gap between physical and digital information spaces while seeking information?

– which factors affect the users' intention to adopt such a technology, and in which ways? Which are the dominant?

Technology acceptance theories provide powerful decision models and reasoning tools for the investigation of the posed questions. The most recent is the Unified Theory of Acceptance and Use of Technology (UTAUT), presented by Venkatesh et al. [14] and integrates successfully the most popular and commonly used technology acceptance models. According to this theory four factors were found to have a direct significant effect on intention to use and actual use behavior: performance expectancy (PE), effort expectancy (EE), social influence (SI) and facilitating conditions (FC). *Intention to use (or Behavioral intention, BI)* describes an individual's readiness to display a certain behavior, i.e., use a given technology to accomplish certain tasks, while *use behavior* (UB) describes the extent to which an individual really uses a given technology. *Performance expectancy* (PE) is defined as the degree to which an individual believes that using a given technology will help him/her to attain gains in job performance. According to UTAUT model there are five factors directly affecting performance expectancy and these are: (a) *perceived usefulness* (PU), the degree to which a person believes that using a particular system would enhance his or her job performance, (b) *extrinsic motivation* (EM) that describes the perception that users will want to perform an activity because it is perceived to be instrumental in achieving valued outcomes that are distinct from the activity itself, such as improved job performance, pay or promotion (c) *job-fit* (JF), that describes the degree to which individuals find that the capabilities of a system enhance their job performance, (d) *relative advantage* (RA), the degree to which individuals believe that using an innovation is perceived as better than using its precursor and (e) *outcome expectations* (OE), that refers to personal expectations regarding the individual's goals and it is the degree to which an individual believes that using a system will lead to better results, such as improvement in work quality and will subsequently gain recognition by other individuals. *Effort expectancy* (EE) is defined as the degree of ease associated with the use of a given technology. On the other hand, effort expectancy is known to be directly affected by three factors: (a) *perceived ease of use* (PEoU), the degree to which a person believes that using a system would be free of effort, (b) *complexity* (CO), the degree to which a system is perceived as relatively difficult to understand and use and (c) *actual ease of use* (EoU), the degree to which using an innovation is perceived as being difficult to use. *Social Influence* (SI) is defined as the degree to which an individual perceives that other individuals who are important to him/her believe that (s)he should use the given technology. According to UTAUT it directly affects the intention to use a given technology. This factor is represented by (a) *subjective norm*, which is defined as the individual's perception that most people who are important to him think (s)he should (or should not) display a certain behavior, (b) *social factors*, which are individual's internalization of the reference group's subjective culture, and specific interpersonal agreements that the individual has made with others, in specific social situations, (c) *image*, the degree to which use of an innovation is perceived to enhance one's image or

status in one's social system and (d) *social norm. Facilitating conditions* (FC) are defined as the degree to which an individual believes that an organizational and technical infrastructure exists to support use a given technology. This factor is the resultant from three factors: (a) *perceived behavioral control*, the degree of perceived ease (or difficulty) of displaying a certain behavior, (b) *facilitating conditions*, which are the set of objective factors in the environment that observers agree that they make an act easy to accomplish and (c) *compatibility*, which is the degree to which an innovation is perceived as being consistent with the existing values, needs, and past experiences of potential adopters. Besides the factors that directly influence intention to use a system, self-efficacy, anxiety and attitude towards using technology are theorized not to be direct determinants of intention. *Attitude toward using technology* (ATUT) is defined as an individual's overall affective reaction to using a system and describes an individual's liking, enjoyment, joy and pleasure associated with use of a given technology. UTAUT was quickly acknowledged and adopted due to its superior interpretive power in mobile information systems literature. For instance Theng et al. [12] used UTAUT to explore the acceptance of mobile technology in digital libraries in the field of geography, for pedagogical reasons and found that the most important factors were perceived usefulness, perceived ease of use and attitude towards use.

In the current context of study, there are some limitations that call for the exclusion of certain factors in the UTAUT model; extrinsic motivation (EM) is excluded because the library patrons who use the new technology will not attain any pay increment or promotion for using it. In addition, social influence (SI), facilitating conditions (FC) and actual use behavior (UB) are meaningful only when the technology has been used for adequate time, to build and establish the corresponding trends, infrastructures and behaviors. In this study we investigated the factors affecting the users' intention to use the mobile technology in a hybrid information environment by evaluating their perceptions during interaction and use with a prototype system over a short period, just developed for that purpose. Performance expectancy (PE) is expressed as the combined effect of perceived usefulness (PU), outcome expectations (OE), job fitness (of the technology used) (JF) and attitude towards using the technology (ATUT). Effort expectancy (EE) is expressed by perceived ease of use of the mobile computing technology (PEoU). The adopted model, shown in Fig. 1, consists of the factors under investigation and the interactions between them. The interaction causality between factors is represented by the paths connecting the factors, leading to certain hypotheses:

- H1: Performance expectancy has an effect on behavioral intention;
- H2: Perceived usefulness has an effect on behavioral intention via performance expectancy;
- H3: Perceived ease of use has an effect on behavioral intention via perceived usefulness and performance expectancy.

We followed an iterative process of system design, starting from the identification of exemplary practices in realistic contexts and continued with a focus group

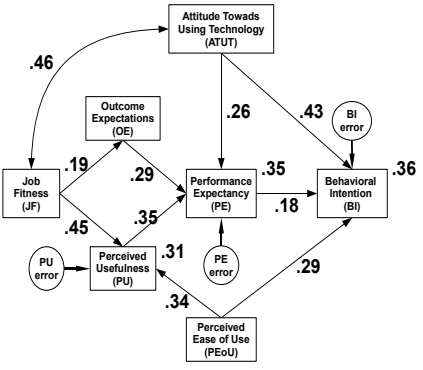


Fig. 1. The proposed model and the validation results

Fig. 2. A snapshot from the dbWiz search form

with information and computer researchers to identify the functionalities and services to be provided. In the next phase, we developed a prototype system which would support information seekers in exploring the available physical and digital collections from anyplace within a library's setting. A pilot study was conducted [15] and subsequently we proceeded (after proper modifications) to the main, quantitative approach for the evaluation of the prototype system and the validation of the model presented in Fig. 1. This mixed methods approach allowed us to collect valuable information about the impact of mobile devices in hybrid information environments on search tasks, identify flaws in the prototype and test the data collection instruments and tools, before proceeding to a big scale field study.

4 The Prototype

The prototype system aimed at supporting its users by providing a collection of tools and corresponding functionalities for their most frequently used information seeking practices, i.e., searching, browsing and asking for advice or further information. These were: (a) search the available sources: we used a meta-search engine[1] to search through the available printed and digital collections available, suitable for the small screen size of a mobile device. The user submits his query once and collects the results from multiple sources (e.g., the local library catalog, ACM, Science Direct, IEEE and Google). The search form is shown in Fig. 2. WiFi connection allows for search and retrieval from anyplace within a library; (b) provide map-based guidance to books: the user spots the DDC[2] number of

[1] dBwiz: `http://researcher.sfu.ca/dbwiz`
[2] Dewey Decimal Classification.

a book of interest and follows a link that leads to a floor-plan, indicating the location of the book. (c) provide a look-up table for the DDC scheme: the user can find the subject of a book, just by knowing its DDC number. He can then submit a query for other books in similar subjects. (d) exchange short messages with other users: this allows the user to send and receive short information abstracts via email or other applications to themselves, friends or library staff; (e) download, share and synchronize digital content with user's personal computer: the mobile device is used as an information harvesting tool. Having collected the desired documents and information the user can transfer it to a desktop computer where he can easily study it; (f) easily create notes: the users can keep some notes, usually about book metadata while being on the stacks. These notes can later be reviewed, shared or transferred. These functionalities were selected after the results of the focus group and the pilot study. During the focus group study it was identified that the indoor environment raises some technical and policy problems, such as authenticated access, location awareness and privacy considerations respectively. The latter was confirmed by the users during the pilot study and therefore it was preferred to follow a context-sensitive instead of a context-aware approach. Furthermore, due to these preferred functionalities we declined the option to generate user tasks that involved interactions with sensor points, such as RFID tags and QR Codes that had been used in pilot tests. In a parallel study that took place in the same setting it was found that QR Codes, though useful and convenient for the organization, such as being low cost, having increased data storage, etc., were found by the users to be cumbersome and to cause problems that slowed the seeking process [11]. On the other hand, the lack of universal standards on RFID technology and the large size of the antenna required has led most manufacturers of mobile computing devices to omit this feature from their devices. Within the library environment, this technology is used in the modern librarian's inventory for certain tasks, such as book lending transactions.

5 Survey Design and Findings

5.1 Data Collection

To validate the evaluation model shown in Fig. 1 we recorded the students' perceptions towards the evaluation criteria presented in the model, by means of a self-reported, multi-scale, multi-item questionnaire, which had been previously tested by the developers of the UTAUT model [13]. Also, usability of a given technology is known to affect its perceived usefulness (PU) and the intention to use (BI), as indicated in the adopted model. Therefore, we used a usability questionnaire[3] to assess ease of use of the prototype system and identify the factors that mostly affect the overall usability and subsequently the BI factor. The usability questionnaire was based on the structure of the QUIS questionnaire [6] to which we added some items regarding text input using a stylus and a virtual

[3] Available at `http://dlib.ionio.gr/hls/en/patras.html`

keyboard resulting in 29 self-reported, questionnaire items to be assessed in a 0 to 6 scale.

The survey was conducted in the premises of the Central Library at Patras University. Students from faculties of engineering, social and health sciences were invited to participate and were offered a small reward as motivation. In four weeks we had 111 students registered: 28.6% were female and 71.4% were male; 81% were undergraduate and 19% were post-graduate students; 90% came from science and engineering faculties whereas 10% came from humanities/health faculties. A PDA was given to each participant to perform tasks that conformed to the following generic scenario: "*a student uses the on-line meta-search engine to submit a query to the available sources. She finds a document of interest from the printed collection and uses the on-line map tool to locate in the stacks. She spots the book, browses through its content, takes a look at nearby documents and takes some notes. While being on the stacks, she browses through the search results from the electronic sources and by reading their abstracts she evaluates their relevance to the books in front of her. After downloading the full-texts she forwards by email to a personal account the collected information* ". Subsequently, the users were asked to express how they perceived the factors of the model by filling the UTAUT questionnaire. The students were also invited to assess the ease of use of the prototype system by filling the usability questionnaire.

5.2 Data Analysis

The interactions among the factors described in Fig. 1 were analyzed using AMOS (a path analysis software). The numbers over the paths are proportional to the relative size effect of the interactions, they were found to be statistically significant and the recorded data fit the model adequately (χ^2(df=11)= 16.31, p>0.05, GFI= 0.96). The numbers in the upper-right corner of the predicted variables BI, PE, PU and OE represent the variance explained by their predictors. The path model depicts how the interactions among factors are mediated to intention to use. The curved arrow connecting JF and ATUT indicates that the two factors are related, probably by means of a mediated factor not taken into account in the current model. The straight arrows represent direct effects. For instance, perceived ease of use has a direct effect on intention to use the mobile technology (0.29) and a mediated effect via perceived usefulness and performance expectancy. As shown in Table 1 the total effect on use intention (BI) is 0.309 which means that by improving the perceived ease of use score by one standard deviation (of the PEoU variable) the intention to use will increase by 0.309 standard deviations[4], i.e., 0.309×1.05= 0.325 units. Both attitude towards using technology and perceived ease of use were also found to have moderate, direct effects on intention to use. In general the model explains 36% of the variance in the users' intention to use a mobile device while exploring hybrid information spaces. This is lower than the interpretive power of the full UTAUT model but considering the factors taken into account (social influence and facilitating

[4] The standard deviation of BI was found to be 1.05.

Table 1. Total effects among model factors

	ATUT	PEoU	PE	PU	OE	JF
PU	–	.341	–	–	–	.446
PE	.257	.120	–	.353	.294	.214
BI	.472	.309	.181	.064	.053	.039

All effects were found statistically significant

conditions were excluded) and since it lies between 0.20 and 0.50 it indicates a moderate to strong interpretive power of the model [1].

Regarding the usability analysis we collected 83 questionnaires; the reliability of the questionnaire was indicated by Cronbach α, which was found above the acceptable threshold of 0.70 (α_{PU}=0.89, α_{PEoU}=0.84, α_{BI}=0.96). The overall usability score was found to be 4.32 (STD= 0.58). In the upper part of Table 2 we present the average scores of overall reactions towards usability of the prototype which was found to be an interesting and useful tool for the unified exploration of hybrid information spaces. The supported functionalities have been perceived as effective and adequate, thus enhancing seeking performance which is known to affect intention to use, as shown in Fig. 1. The mobile device given was found easy to use and overall, the prototype use has been satisfactory. We have also calculated the average scores across five factors known to affect usability: presentation, text input, terminology and language, learnability and system capabilities. The scores indicate adequate usability, with the highest values recorded on easiness to learn how to operate the prototype (AVG=4.64, STD= 0.72) and the lowest value recorded in text-input process (AVG=3.72, STD= 1.0), which was a new experience for many participants.

To gain a better insight about the factors affecting the system's usability we conducted group comparisons. Within-subject analysis (t-tests) detects differences among responses of the users themselves. It revealed the two sets of user responses at questionnaire statements that mostly deviate (positively or negatively) from the total usability average score. The t-values shown in the lower part of Table 2 are analog to the differences between mean scores. Between-subjects tested for a significant difference in overall usability score among students from science/engineering departments (group0) versus students from humanities and health sciences (group1) and the difference was not found to be significant (t(df=82)=-1.13, p=0.261). The average overall usability scores were found to be $score_{gr0}$= 4.3 (STD= 0.56) and $score_{gr1}$= 4.54 (STD= 0.73). ANOVA analysis was also used to explore the change of overall usability score across four levels of experience in the use of mobile devices such as PDAs, i.e., none, some, medium, high. We hypothesized that the bigger the experience the greater the overall usability score. Three contrasts were tested; none versus some, some versus medium and medium versus high experience. The average overall usability values for each level of experience were found to be respectively 4.14 (STD= 0.56), 4.24 (STD= 0.70), 4.51 (STD= 0.45) and 4.51 (STD= 0.43).

Table 2. Average scores of usability questions

Average usability score: **AVG= 4.32, STD= 0.58**					
Overall reactions	*AVG*	*STD*	Prototype characteristics	*AVG*	*STD*
Easy	4.25	1.16	Learnability	4.67	0.72
Satisfactory	4.55	0.94	Terms & Language	4.37	0.77
Adequate	4.63	0.88	Capabilities	4.23	0.73
Effective	4.62	0.86	Presentation	4.12	0.84
Interesting	4.70	1.08	Text input	3.72	1.0
Useful	4.85	0.88			

Top positively deviated items	t-value	STD
1. Learnability: error reduction	12.08	0.95
2. Learnability: memory load	7.62	0.63
3. Effectiveness: search results' relevance	7.07	0.58
4. Useful: usefulness	5.90	0.52
5. Interest: service perceived as interesting	4.13	0.58
Top negatively deviated items		
1. Text input: typing minimization	-7.76	-1.22
2. Presentation: screen navigation	-6.44	-0.85
3. Text input: typing speed	-4.60	-0.53
4. Text input: self error correction	-3.35	-0.38
5. Capabilities: response time	-3.20	-0.34

5.3 Discussion

The students' average score on intention to use (BI) the mobile technology was found 4.33 (STD= 1.05) and more than 80% and 70% of records on effectiveness (PE) and usefulness (PU) respectively were assessed over 4.0, showing that the users perceived the mobile technology as capable of enhancing their seeking experience while exploring the hybrid information space. As shown in Table 1 the strongest influence on intention to use (BI) comes from the attitude of users (ATUT) towards using a mobile device, i.e., their overall liking and enjoyment associated with the use of a small size computer. Provided they are not prejudiced against the use of the mobile technology, we found that perceived ease of use (PEoU) has a stronger effect on BI, compared to performance expectancy (PE). This means that the students are willing to explore and adopt the proposed seeking approach, overlooking the performance expectancy, as long as it is easy to use.

The results from usability analysis presented in Table 2 indicate that the users found the given prototype system easy to use. Within-subjects analysis revealed that error reduction and low memory load were rated quite higher than the overall average, showing that prototype's learnability did not puzzled the users and therefore increased perceived ease of use and subsequently intention to use and adopt the mobile technology. Similarly, factors affecting performance expectancy (relevance and usefulness) were also recorded higher than the average usability score. Also, the users perceived the new service as interesting denoting

a clear attitude towards using the new technology. On the other hand, lower than the overall average were the scores on text input (in terms of typing minimization and speed), probably due to lack of word automatic completion and drop-down lists which reduce typing errors and increase speed of text input. The lower than the average score for screen navigation indicates that users need more visual navigation cues, such as sitemap trees. This was due to the small screen size of the device which leads in a progressive disclosure of the information content. The between-subjects analysis showed that the overall usability score from the Science/Engineering group did not significantly differ from the group of students coming from other faculties, i.e., prior familiarity and attitude towards mobile computing did not seem to affect perceived usability. Therefore the hypothesis that "*the biggest the level of experience the greater the usability*" can be rejected. In addition, the usability score for all groups was above 4 which indicates that all the students found the prototype easy to use.

Regarding the prototype's functionalities, the study revealed that mobile computing devices can be mostly used: (a) to access on-line search tools, (b) to provide anywhere availability of reference tools and (c) to exchange short messages with other users with communication tools. Students were asked to rank the supported functionalities, by assigning to each one an integer between 1 (least useful) and 7 (most useful). The average score for these functionalities was found to be 5.33 (SE= 0.22) for the federated search engine, 4.31 (SE= 0.18) for the map guidance, 3.69 (SE= 0.17) for the e-mail application and 3.4 (SE= 0.19) for the instant messaging application.

6 Conclusions

The analysis allowed us to rank the factors affecting the intention to use the proposed system, according to their effect size. We have found that attitude towards mobile computing has the strongest effect in behavioral intention and it is more than double of the effect of performance expectancy. Perceived ease of use was also found to have a significant effect on intention to use the mobile technology. We conclude that provided users are not prejudiced against mobile devices and that they find easy to use, they are willing to adopt the proposed seeking approach for the exploration of hybrid spaces, considering its effectiveness a lower priority. This indicates the users' need for new information services and tools to explore the modern information landscape. Usability analysis revealed that students found the prototype system usable and easy to learn and therefore more likely to adopt a system based on mobile computing. It also showed us which are the weak points of the prototype, so that further improvements can be made. In addition, by ranking the supported functionalities we found that students prefer to use the computing capabilities of the mobile device for information harvesting rather than storing information for later reference.

For the near future, we plan on recording the actual use behavior of the prototype system as well as the factors affecting it and subsequently compare its use to the currently available seeking system. In addition, we plan to assess the

impact of additional information retrieval tools, based on image processing from camera enabled devices and recommendation algorithms that enhance data and information extraction from large volumes of available resources.

References

1. Acock, A.C.: A Gentle Introduction to Stata, 2nd edn. Stata Press (2008)
2. Aittola, M., Ryhänen, T., Ojala, T.: Smartlibrary – location-aware mobile library service. In: Chittaro, L. (ed.) Mobile HCI 2003. LNCS, vol. 2795, pp. 411–416. Springer, Heidelberg (2003)
3. Buchanan, G., Pearson, J.: An architecture for supporting rfid-enhanced interactions in digital libraries. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 92–103. Springer, Heidelberg (2010)
4. Buchanan, G.R.: The fused library: integrating digital and physical libraries with location-aware sensors. In: Proc. of 10th JCDL, pp. 273–282. ACM, New York (2010)
5. Jones, M.L.W., Rieger, R.H., Treadwell, P., Gay, G.K.: Live from the stacks: user feedback on mobile computers and wireless tools for library patrons. In: Proc. of the 5th ACM Conf. on Digital libraries, pp. 95–102. ACM, New York (2000)
6. Lin, H.X., Choong, Y.Y., Salvendy, G.: A proposed index of usability: a method for comparing the relative usability of different software systems. Behavior and Information Technology 16(4-5), 267–277 (1997)
7. Lin, H.T., Lin, C.F., Yuan, S.M.: Using RFID guiding systems to enhance user experience. The Electronic Library 27(2), 319–330 (2009)
8. Mäkelä, K., Belt, S., Greenblatt, D., Häkkilä, J.: Mobile interaction with visual and rfid tags: a field study on user perceptions. In: Proc. of Conf. on Human Factors in Computing Systems, pp. 991–994. ACM, New York (2007)
9. Ryan, C., Gonsalves, A.: The effect of context and application type on mobile usability: an empirical study. In: Proc. of 28th ACSC, pp. 115–124. Australian Computer Society, Inc. (2005)
10. Satpathy, L., Mathew, A.P.: Rfid assistance system for faster book search in public libraries. In: CHI 2006 extended abstracts on Human factors in computing systems, CHI EA 2006, pp. 1289–1294. ACM, New York (2006)
11. Stoica, G.A.: An architecture to support context-aware mobile applications. Phd, University of Patras (2010)
12. Theng, Y.L., Tan, K.L., Lim, E.P., Zhang, J., Goh, D.H.L., Chatterjea, K., Chang, C.H., Sun, A., Yu, H., Dang, N.H., Li, Y., Vo, M.C.: Mobile G-portal supporting collaborative sharing and learning in geography fieldwork: an empirical study. In: Proc. of 7th JCDL, pp. 462–471. ACM, New York (2007)
13. Venkatesh, V., Davis, F.D.: A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science 46(2), 189–207 (2000)
14. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: Toward a unified view. MIS Quarterly 27(3), 425–478 (2003)
15. Veronikis, S., Gavrilis, D., Zoutsou, K., Papatheodorou, C.: Using handhelds to search in physical and digital information spaces. In: Proc. of 2nd Int. Conf. on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp. 225–230. IEEE CS, Los Alamitos (2008)

Digital Library 2.0 for Educational Resources

Monika Akbar[1], Weiguo Fan[1], Clifford A. Shaffer[1], Yinlin Chen[1], Lillian Cassel[2], Lois Delcambre[3], Daniel D. Garcia[4], Gregory W. Hislop[5], Frank Shipman[6], Richard Furuta[6], B. Stephen Carpenter II[7], Haowei Hsieh[8], Bob Siegfried[2], and Edward A. Fox[1]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
[2]Department of Computing Sciences, Villanova University, Villanova, PA, USA
[3]Department of Computer Science, Portland State University, Portland, OR, USA
[4]Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA, USA
[5]Department of Computer Science, Drexel University, Philadelphia, PA, USA
[6]Department of Computer Science, Texas A&M, College Station, TX, USA
[7]Art Education Program, Pennsylvania State University, State College, PA, USA
[8]School of Library & Information Science, Univ. of Iowa, Iowa City, IA, USA
{amonika,wfan,shaffer,ylchen}@vt.edu, lillian.cassel@villanova.edu, lmd@cs.pdx.edu, ddgarcia@cs.berkeley.edu, hislop@drexel.edu, {shipman,furuta}@cs.tamu.edu, bsc5@psu.edu, haowei-hsieh@uiowa.edu, rsieg@ptd.net, fox@vt.edu

Abstract. We report on focus group feedback regarding the services provided by existing education-related Digital Libraries (DL). Participants provided insight into how they seek educational resources online, and what they perceive to be the shortcomings of existing educational DLs. Along with useful content, social interactions were viewed as important supplements for educational DLs. Such interactions lead to both an online community and new forms of content such as reviews and ratings. Based on our analysis of the focus group feedback, we propose DL 2.0, the next generation of digital library, which integrates social knowledge with DL content.

Keywords: Digital Library 2.0, Computing Portal, Ensemble.

1 Introduction

The information needs of digital library (DL) audiences vary widely depending on the nature of the digital library. In this paper we focus on the needs of educators for teaching and learning. We seek to provide a digital library that supports the communities of educators, because in real-life, educators often share their resources and experiences with each other. Ensemble[1], the computing education portal within the National Science Digital Library (NSDL), supports a wide range of computing education communities, provides resources for developing programs that blend computing with other STEM areas (e.g., *X-informatics*

[1] http://www.computingportal.org/

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 89–100, 2011.

and *Computing+X*), and seeks to produce digital library innovations that can be propagated to other NSDL pathways. Ensemble is a distributed portal providing access to the broad range of existing educational resources while preserving the identity of the individual collections and their associated curation practices. Ensemble encourages contribution, use, reuse, review, and evaluation of educational materials at multiple levels of granularity.

Ensemble made public the beta version of its web site in March 2010 and is now working on launching the production version. As the project moved forward, researchers at Virginia Tech conducted two focus groups comprised of nine business faculty members who teach computing to business majors. These participants constitute the majority (90%) of the department of Business and Information Technology (BIT). This pool broadens our perspective as it has different information needs compared to Computer Science educators — a group which dominates the Ensemble project team.

We anticipated learning about the techniques and challenges educators face when using online resources. Analysis of the discussions indicates that the problems we face in serving these users are related to the quality and quantity of content as well as the ability to manage those contents. We also found that current DLs can be improved if they support social interactions. Based on our findings, in this paper we propose DL 2.0, which integrates user interactions with resources in a DL. Thus our findings have the potential to be useful across various education communities as well as other digital libraries.

2 Prior Work

There have been a number of efforts to define a digital library [10,14]. Quality evaluation for DLs also has seen a fair amount of research [6,8,24]. Many of these articles pointed out the importance of understanding the needs of the target audience. Xie [29] identified major areas that contribute to the success of a DL: usability, quality of collection, service, and system performance. All of these are building blocks of a successful information system [5,25]. Researchers have pointed out different aspects of establishing an online community in a DL [4,19,12]. There has been significant research on design issues [2,11], studying and analyzing the overall architecture [26,28], and identifying the success factors [15,16] of online communities.

Online communities depend on user interaction to become active and stay useful. Girgensohn, et al. [7] identified three sociological design challenges for building a successful socio-technical site: encouraging user participation, fostering social interactions, and promoting visibility of people and their activities. Koh, et al. [13] noted that participation can be of two types: passive participation (i.e., viewing) and active participation (i.e., posting), and each of these activities depends on different stimuli which include active leadership, offline interaction, content usefulness, and sound infrastructure. User participation in online communities has been studied in depth from various angles. Nov, et al. [21] studied various motivations for different types of participation for varying

levels of membership in the community. Luford, et al. [17] studied the effect of showing both similarity and distinctness information about a member and the groups where he or she belongs as a means for increasing online community participation. Beenen, et al. [3] did similar studies based on social theories. Millen, et al. [20] investigated factors such as design decisions, member selection, and facilitating stimulating discussion as means of engaging the members of an online community. Preece, et al. [23] studied community members to find out reasons behind lower participation rates of a particular group of less active users known as *lurkers*.

While prior research focused either on the success of DLs or on the success of online communities, we are proposing a combination of these areas to drive the future of Digital Libraries.

3 Data Collection and Analysis

There were two main phases in our research as described in Table 1: data collection and analysis. The department of Business Information Technology (BIT) at Virginia Tech has a unique pool of computing educators who teach IT and CS courses to Business majors. We invited 10 faculty from this department. Five were present at the first session, and four at the second. Each session was an hour long.

Table 1. Phases of Data Collection and Analysis

Data Collection	
System Review	Identified key areas of Ensemble for further research and development.
Protocol Dev.	Created a protocol and a set of questions for the focus groups.
Focus Groups	Virginia Tech (VT) conducted two focus groups. Each focus group was roughly one hour in duration.
Participants	Each of the 9 participants were Business faculty who teach computing to Business majors.
Data Analysis	
Transcription	Audio recordings were transcribed and combined with handwritten notes taken during the session to create a combined report of the two focus groups.
Coding	We identified repeated answers, patterns, and behaviors in the transcribed data and in the report. These were coded based on the themes they represented.
Themes	The codes were used to identify emerging themes which were then used to develop and connect high-level codes about the prevalent practices on locating and using electronic resources, on creating active users in an educational DL.

Our questions to participants were split across two broad topics: (i) How do they search for educational materials? and (ii) What is their feedback on the Ensemble portal? We posed a set of 10 questions based on these two broad topics, which are listed below, to all participants.

1. How do you search for resources to use in a course, lesson, or assignment related to an IS/IT-oriented course?
2. In which content areas would you normally seek resources to support learning and teaching?
3. Which formats might be most helpful to your teaching or your students' learning?
4. Which resources do you have the most difficulty finding and accessing?
5. How do you stay up-to-date in your field in terms of education?
6. Which web sites do you visit or which materials do you make regular use of? Why?
7. Do you use publisher sites often for your assessment needs?
8. Do you participate in any special interest groups (SIGs) or meetings to enrich your teaching or any social group? Do they have an online community site for it?
9. How valuable do you consider the use of badges and rewards in building an online community?
10. What are your thoughts about the Ensemble web site?

While in this paper we report and use the data from two focus groups consisting of faculty members, similar studies were done by members of our team at a variety of locations including the University of Iowa, where 25 students from the Library and Information Science department participated in five focus groups. Results from those focus groups identified most of the issues addressed

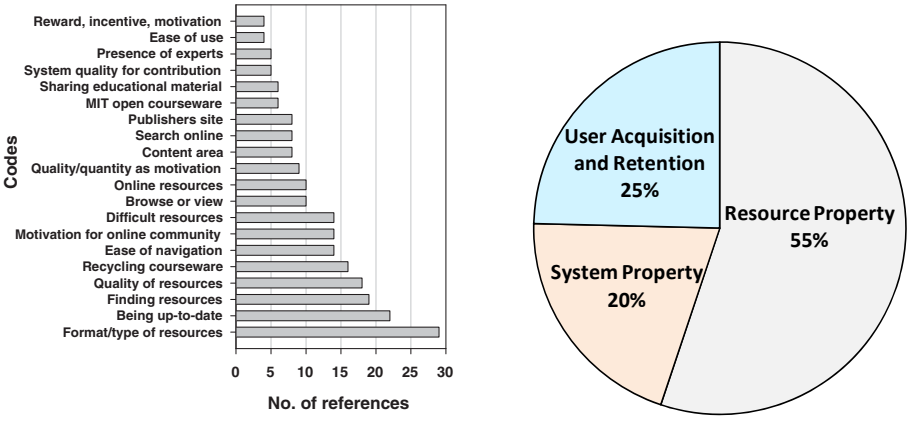


Fig. 1. (Left) Sample codes with number of references. (Right) Distribution of references in three major themes described in Table 2.

.

by our faculty participants, though students were less concerned with finding and reusing course-material. They were interested in course materials that were printer-friendly. We found no significant difference between the two groups, i.e., educators and students. In this paper, to be precise, in subsequent discussion we only use data from Virginia Tech faculty participants.

We followed the grounded theory approach [27] to analyze the data. Initial coding was done to identify recurring themes or examples related to a theme which resulted in 29 codes. Many of these codes relate to an underlying broader theme which helped us to identify different aspects of the code. For example, the code *Ease of navigation* (14 references) referred to various aspects of navigating through a site. While some participants argued that *organization* of content is a major issue for easy navigation, others were inclined toward better *search mechanisms*. We did not tie specific codes to specific questions. Participants provided more information as we progressed through the sessions, causing the same code to be linked with multiple questions. There were 246 references to these codes in the original transcripts. Figure 1(a) shows some of the top codes with their reference counts. For example, participants mentioned *format or type of the resources* 29 times. YouTube and educational video clips were mentioned as either motivating tools for students or informative resources. There were also mentions of syllabi, lecture notes, and PowerPoint slides that educators often seek on the Internet. *Quality* of available material is also a big concern (18 references). Many participants pointed out that they reuse or borrow existing course material as a starting point (*Recycling courseware*, 16 references).

After the initial coding, we grouped the codes based on their relevance to a set of broader themes. Three themes that emerged were Resource property, System property, and User acquisition and retention (Figure 1(b) and Table 2). Resource property includes types of resources used by educators, difficult resources, methods on how to find resources online, etc. System property lists various aspects of a site that encourage participants to use the site. User acquisition and retention refers to factors that motivate users to actively use a site and participate. Some of the initial codes related to each of these themes are listed below them (see Table 2).

The codes in Table 2 reflect characteristics of an ideal DL, which are similar to those of Web 2.0 [22]. Web 2.0 provides a dynamic environment for users by supporting sets of activities that promote social interactions, encourage user contribution, or capture and highlight collective knowledge. Usefulness of Web 2.0 has been studied for different domains [1,18]. We propose Digital Library 2.0 for educational resources that takes a user-centric approach by providing services to connect users and resources, and by hosting online communities.

4 Resource, Service, and User: Digital Library 2.0

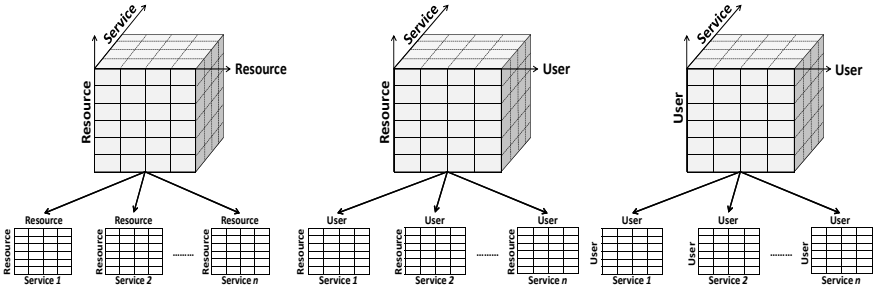
Our focus group sessions uncovered a series of unmet needs for educational resources, which include a digital library with rich resources, dynamic interactions between users and resources, and an active virtual community.

Table 2. Emerging Themes from the Focus Group Data

Resource Property
Format: Types/formats of educational materials.
Finding resource: Finding resource through Web search (e.g., Google), university sites (e.g., MIT OpenCourseWare), and personal connection.
Quality: Quality of available resources at various sites.
Recycling courseware: Reusing course material or borrowing course content.
System Property
Factors influencing site use.
Ease of navigation - Organization of content: Easier topical organization following any standard organization scheme.
Robust search: Visible search box/tab and granular searching options.
Interface: Takes less time to get used to and use the resource.
Association between content.
Factors influencing contribution.
Ease of contribution: Contribution should not take time.
Personalization
Content customization: Ability to customize textbook or assessments.
Add content to user list: Create personal collection from existing resources.
Differential access to resources: Access control to resources, especially for assessment materials.
User Acquisition and Retention
Motivation for using the site.
Existence of quality resource.
Existence of large quantity resource.
Existence of peer reviews.
Existence of experts in the community.
Critical mass: Large user base.
Saving time as a motivation for joining an educational DL.
Motivation for contribution
Peer recognition.
Quality of community and resources in the site.
Reward, incentive.
Academic recognition for contribution (e.g., Promotion and tenure).
Building reputation (e.g., roles, badges) based on user activities.
Peer recognition.

Participants mentioned a number of services they would like to see which relate *resource* and *user*. Different connections between and among *resource* and *user* can create different relationships between these types of entities that can provide better exposure of resources and can eventually lead to better use of content. In some cases, these relationships can even produce new content. For example, services that connect a user with resources might allow the user to generate new content in the form of ratings or reviews.

Formal Definition: DL 2.0 is a combination of three basic entities: $\mathcal{R}$, S, and U (resource, service, and user, respectively). DL 2.0 architecture is dependent on



(a) RSR Relational Matrix (b) RSU Relational Matrix (c) USU Relational Matrix

Example of services (in bold text) for each relational matrix

- Linking resources (e.g., tags).
- Associating resources (e.g., exercises linked to a lecture slide).
- Peer reviews (e.g., ratings).
- A resource can have an **owner**.
- A resource can be **read**/**downloaded**.
- Users can **contribute** additional information (e.g., comments, ratings).
- Users can be **members** of a group or community.
- Users can **contact** other users.
- Users can be **connected via shared resources** (e.g., co-authors).

Fig. 2. Relationships between Resource and User

three different arrangements of the basic entities: $\{\mathcal{RSR}\}$, $\{\mathcal{RSU}\}$, and $\{\mathcal{USU}\}$. It indicates that service is the connecting entity in relating resources with other resources, resources with users, and users with users.

A service that connects two entities can implicitly create connections between or among other entities. Figure 2 shows these relations with examples. Figure 2(a) shows the Resource-Service-Resource (RSR) relational matrix. For each service in this relational matrix, there will be relationships between some of the resources. For example, a resource might contain annotations (which is another type of resource). Figure 2(b) shows the Resource-Service-User (RSU) relational matrix. A resource can be connected to users via a number of services such as author or viewer. Figure 2(c) presents the User-Service-User (USU) relational matrix. Connections and interactions between users would allow for a virtual social environment that is desired by a large number of participants.

4.1 Resource-Service-Resource (RSR) Relational Matrix

More than half of the codes from our initial data analysis phase were related to some property of resources (see Figure 1, right). **Organization** and interconnection between resources are important to users. Participants identified a number of problems with various organization schemes used at different sites, with the most common being learning the many different organization schemes. One

suggestion was to use existing standards to create the categorization scheme. This would allow all resources to be organized by a set of well-known topics. Use of non-standard terms was also confusing to many users. **Association** between content can be useful to users. Participants noted that they like to explore and use resources that are related to their course content. This highlights the fact that an individual resource page serves few information needs of educators who would prefer the resources to be linked properly. DLs need to have a robust organization scheme of content and proper association between various resources.

Approachable **navigation** is important for encouraging users to explore a DL. Using deep navigation trees can be confusing. If the content is buried under five or six levels, a user often loses track of the context. Tags or lists with low depth can be useful. One suggestion was to show the context (e.g., tree, bread crumb). When applicable, information such as the link to the actual content should be *eye-catching* or visually appealing. It was suggested that for a DL that hosts groups and communities, the navigation scheme should be consistent across collections, communities, and other sections. **Search** is considered as an essential service. Several participants mentioned frequent use of advanced search features to locate relevant materials among a large number of resources. This feature is used even by those who are familiar with the site.

Quantity and **quality** of content is another recurring code. Aside from the services related to resources (e.g.,personalization), more information on the content was viewed as useful. **Additional information** can come in various forms such as description of the resources, peer reviews, ratings, comments, or usage notes. All of this information requires that there be a "group of users" who "actively participate" in the DL.

4.2 Resource-Service-User (RSU) Relational Matrix

One defining aspect of DL 2.0 is that users will play a key role here. Static resources are not enough to meet many of the information needs of users, especially the educators. There exists a need for a system that would allow educators to interact with the resources and contribute easily. Systems that have peer reviews were appreciated by the participants. Such reviews can appear in various forms and require that a system is flexible enough to include those services whenever needed. Above all, **ease of contribution** is critical in the success of DL 2.0.

One of the prevalent practices among educators is recycling courseware. Depending on the audience and the syllabus, they may reuse some of the course materials or introduce new content. Thus, having the ability to customize the content to fit the demands of a course can be crucial to educators.

Usability is another issue for the next generation of DL. While users like more information, they also tend to prefer a clean interface. When the site contains much information, the *search* option is rapidly sought out by users. Getting used to the site should not take much time, as one participant explained, "it is unlikely that someone would spend too much time to figure out how it can be used." Time is a scarce resource for educators. They want a system that lowers their prep time, not one that requires time to understand.

One way that we can help users save time is by introducing personalization features such as annotation or ability to tag content. Notifications can help users stay connected with the site. Several participants mentioned subscribing to news-feeds. Being notified about chosen content or users is a form of personalization that can help the users stay connected while not taking too much time.

4.3 User-Service-User (USU) Relational Matrix

Community feedback and peer reviews are important when trying to locate and use quality educational material. Social interactions in virtual environment can take place in a number of formats including comments, ratings, and tags (CRTs). Various sites depend on forums or blogs to share information on a larger scale. While most of these services create implicit connections between users, there are services that directly link one user with another (e.g., contact forms, message windows, an option to create a personal network). While these options would allow users to communicate with each other and stay connected, we first need to motivate users to visit the site and explore the content.

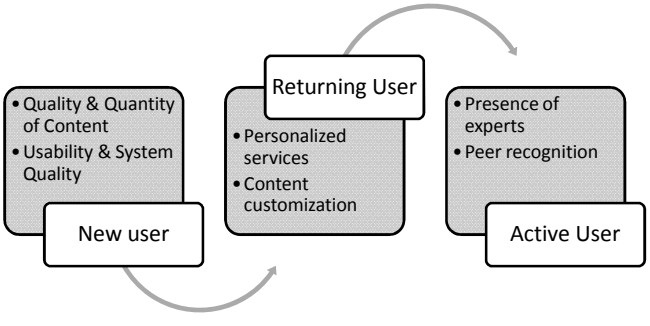


Fig. 3. Types of Users and Motivating Factors

Various factors act as motivator to encourage users to visit an educational resource site, use the materials, and actively participate in the community. We broadly divide users into three categories based on level of activity: new user, returning user, and active user. Each user type needs certain motivations to stay in that level or progress to the next level (see Figure 3). For new users, to be useful, a site has to be easy to get used to (usability), have quality materials (content quality), provide useful services (e.g., advanced search, notifications). Motivations for returning users are different as they want the ability to create and share new content, customize content, or specify differential access to resources (e.g., assessments cannot be viewed by students). Returning users may start actively participating once they become used to the site and see value in contributing. Participants mentioned a number of incentives for motivating users to participate in the community. Of these, the presence of experts and active leadership is critical for a successful community. If contributions in the

Table 3. Comparison between DL 1.0 and DL 2.0 based on 5S Definitions

5S Elements	DL 1.0	DL 2.0
Stream	Metadata of resources only.	Metadata with community-contributed information (e.g., comments, ratings, reviews) on resources.
Structure	Single listing of resources belonging to a particular collection/topic.	Cross-referenced resources across collections and attributes.
Space	Does not handle multiple spaces.	Supports multi-layered resource spaces. These layers can support various space-related entities (e.g., time series, feature spaces).
Society	Does not explicitly support group-oriented tasks.	Supports groups, communities, collaborations as well as individual user tasks.
Scenario	Services include browse, index, and search.	Services include personalization, recommendation, better organization, user-friendly navigation, faceted search, advanced ranking based on popularity, users' comments, ratings, tags (CRTs).

community are widely recognized as having value, then they can be useful for career development. Forms of recognition vary based on the type of user. Experts in a field tend to value professional or academic recognition while novice users are satisfied with peer recognition. Recognition can come in the form of badges or rewards. Sharing usage information for a resource with the contributor, which can be used as an impact factor, can be motivating to contributors.

4.4 DL 1.0 vs. DL 2.0

DL 2.0 is the next-generation approach to DL that blends the traditional digital library contents with user-contributed contents and provides online community support (e.g., relationship management among users and digital contents, such as user interactions, rating, comments, bookmarks, querying, etc.). The core difference between traditional DL 1.0 and DL 2.0 lies in the fact that the latter is more dynamic, user-centric, encourages user contribution, fosters virtual community, and incorporates knowledge with resources. While core services of DL 1.0 were limited to search, browse, and indexing, DL 2.0 encompasses content management, dynamic services such as customization or personalization of content, and a collaborative environment. Table 3 provides a comparison of the 5S elements [9] between DL 1.0 and DL 2.0.

5 Conclusions

We have presented focus group data in an effort to explain online information seeking trends of one group of computing educators. We plan to conduct further studies of educators who teach computing science majors. Collected data indicate educators' desire to see improvement over existing educational digital libraries. Educators, in search of quality education material, tend to borrow, adopt, or reuse those materials in their teaching, learning, and research. Many are willing to contribute their knowledge, provided that contribution is not difficult. They make clear that peer review and user contribution are important in educational DLs just as they have proved important to commercial sites such as Amazon.

Based on these findings we propose DL 2.0 services that tie together users and resources to create meaningful relationships. We believe our data provide useful insights on current resource-seeking and resource-usage trends of educators. This information will be beneficial to those who want to develop the next generation of educational digital libraries.

Acknowledgments. This research is supported by NSF Grants DUE-0840713, 0840715, 0840719, 0840721, 0840668, 0840597, 0836940, and 0937863.

References

1. Alexander, B.: Web 2.0: A New Wave of Innovation for Teaching and Learning?. Educause Review 41(2), 32–44 (2006)
2. Andrews, D.C.: Audience-specific Online Community Design. Communications of The ACM - Supporting Community and Building Social Capital 45(4), 64–68 (2002)
3. Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., Kraut, R.E.: Using Social Psychology to Motivate Contributions to Online Communities. In: Proceedings of the CSCW Conference, pp. 212–221 (2004)
4. Borgman, C.L., Bates, M.J., Bates, M.V., Efthimiadis, E.N., Gilliland-Swetland, A.J., Kafai, Y.B., Leazer, G.H., Maddox, A.B.: Social Aspects of Digital Libraries. Final Report for Invitational workshop held at UCLA, February 15-17 (1996)
5. DeLone, W.H., McLean, E.R.: Information Systems Success: The Quest for the Dependent Variable. Information Systems Research 3(1), 60–95 (1992)
6. Fuhr, N., Hansen, P., Mabe, M., Micsik, A., Sølvberg, I.T.: Digital Libraries: A Generic Classification and Evaluation Scheme. In: Constantopoulos, P., Sølvberg, I. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 187–199. Springer, Heidelberg (2001)
7. Girgensohn, A., Lee, A.: Making Web Sites be Places for Social Interaction. In: Proceedings of the CSCW Conference, pp. 136–145 (2002)
8. Gonçalves, M., Moreira, B., Fox, E., Watson, L.: "What is a good digital library?" A Quality Model for Digital Libraries. Information Processing and Management 43(5), 1416–1437 (2007)
9. Gonçalves, 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 (TOIS) 22(2), 270–312 (2004)
10. Greenstein, D., Thorin, S.: The Digital Library: A Biography. Digital Library Federation Council on Library and Information Resources (2002)

11. Gurzick, D., Lutters, W.G.: Towards a Design Theory for Online Communities. In: Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology (DESRIST), pp. 11:1–11:20 (2009)
12. Huwe, T.: Exploiting Synergies among Digital Repositories, Special Collections, and Online Community. ONLINE (Weston, Conn.) 33(2), 14–19 (2009)
13. Koh, J., Kim, Y.-G., Butler, B., Bock, G.-W.: Encouraging Participation in Virtual Communities. Communications of the ACM 50(2), 68–73 (2007)
14. Lagoze, C., Krafft, D., Payette, S., Jesuroga, S.: What is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL. D-Lib. Magazine 11(11) (November 2005)
15. Leimeister, J.M., Sidiras, P., Krcmar, H.: Success Factors of Virtual Communities from the Perspective of Members and Operators: An Empirical Study. In: HICSS 2004, pp. 2708–2715 (2004)
16. Lin, H.: Determinants of Successful Virtual Communities: Contributions from System Characteristics and Social Factors. Information and Management 45(8), 522–527 (2008)
17. Ludford, P.J., Cosley, D., Frankowski, D., Terveen, L.: Think Different: Increasing Online Community Participation using Uniqueness and Group Dissimilarity. In: Proceedings of the SIGCHI Conference, pp. 631–638 (2004)
18. Maness, J.M.: Library 2.0 Theory: Web 2.0 and its Implications for Libraries. Webology 3(2) (2006)
19. Markland, M.: Technology and People: Some Challenges when Integrating Digital Library Systems into Online Learning Environments. The New Review of Information and Library Research 9(1), 85–96 (2003)
20. Millen, D.R., Patterson, J.F.: Stimulating Social Engagement in a Community Network. In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, pp. 306–313 (2002)
21. Nov, O., Naaman, M., Ye, C.: Analysis of Participation in an Online Photo-sharing Community: A Multidimensional Perspective. Journal of the American Society for Information Science and Technology 61(3), 555–566 (2010)
22. Oreilly, T.: What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Communications & Strategies (1), 17–37 (2007)
23. Preece, J., Nonnecke, B., Andrews, D.: The Top Five Reasons for Lurking: Improving Community Experiences for Everyone. Computers in Human Behavior 20(2), 201–223 (2004)
24. Saracevic, T., Covi, L.: Challenges for Digital Library Evaluation. In: Proceedings of the ASIS Annual Meeting, vol. 37, pp. 341–350 (2000)
25. Seddon, P.B., Staples, S., Patnayakuni, R., Bowtell, M.: Dimensions of Information Systems Success. Communication of the AIS 2(3es) (November 1999)
26. de Souza, C.S., Preece, J.: A framework for Analyzing and Understanding Online Communities. Interacting with Computers 16(3), 579–610 (2004)
27. Strauss, A., Corbin, J.: Basics of Qualitative Research: Grounded Theory Procedures and Techniques, 1st edn. Sage Publications, Thousand Oaks (1990)
28. Toral, S.L., Martinez-Torres, M.R., Barrero, F., Cortes, F.: An Empirical Study of the Driving Forces behind Online Communities. Internet Research 19(4), 378–392 (2009)
29. Xie, H.I.: Users' Evaluation of Digital Libraries (DLs): Their Uses, their Criteria, and their Assessment. Information Processing and Management: an International Journal 44(3), 1346–1373 (2008)

An Approach to Virtual Research Environment User Interfaces Dynamic Construction

Massimiliano Assante, Pasquale Pagano, Leonardo Candela,
Federico De Faveri, and Lucio Lelii

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa - Italy
{assante,pagano,candela,defaveri,lelii}@isti.cnr.it

Abstract. Virtual Research Environments are internet-based working environments tailored to serve needs of diverse and evolving user communities. These environments are oriented to promote new ways of dealing with modern research tasks. Their realization requires user interfaces that are dynamically built to provide their clients with *organised views* on the data and services aggregated to meet specific community needs. This paper presents an approach to the problem of Virtual Research Environment user interfaces dynamic construction. This approach is characterized by user interfaces built through a component-oriented strategy and an heuristic for user interface constituents arrangement on the screen. The implementation and exploitation of the proposed approach in the context of the D4Science-II EU funded project is discussed as well as future plans are presented.

1 Introduction

The fast evolution of technologies is setting up a substantial and extensive modification in the model research activities share, aggregate, diffuse, and manipulate outcomes in digital form. Research areas like environmental monitoring, climate change, humanities and social science are affected by this change [7,4,6,8]. These areas are opening at the new technologies as methods to improve the performance of their researchers.

This is the case where *Virtual Research Environments* (VREs) step in. These are innovative internet-based collaborative environments tailored to support scientists to produce and exchange results with peers across the globe in a cost-efficient way. These environments are expected to provide scientists with the *resources* (including data, instruments, processing power, communication tools, services) they need to accomplish their tasks. The main benefits of VREs are enhancing research collaborations over a distance, exchange of information among researchers, access to skills, knowledge, research data and computational resources situated in remote locations [6].

However, the realization of such innovative environments is challenging and requires the development of new approaches and technologies to cope with the distinguishing characteristics of the scenarios VREs should serve, namely their

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 101–109, 2011.

"*highly evolving nature*" in terms of requirements to be satisfied and, consequently, resources to be made available. In this context User Interfaces (UIs) have a fundamental role since it is through them that VRE clients have a perception of the capabilities and resources the supporting environment can provide them with. Their building can not adopt traditional development approaches and requires new ones that are dynamic and adaptive thus to fit with an operational context where the facilities to be made available are not known a priori, instead they are dynamically acquired and aggregated to meet the community expectations [1].

In this paper an approach to the dynamic creation of user interfaces for VREs is described. This approach is based on UIs built in a component-oriented manner and aims at promoting the *sharing* and *reuse* of UI constituents. Component-oriented approaches are becoming popular in web base interfaces and portals because of their capability to suit a particular task or individual, e.g. *iGoogle* make it possible for every user to built its custom interface by adding and removing "widgets". The proposed approach is based on heuristics to automatically select and arrange the UI constituents needed to satisfy a "VRE specification", i.e. an abstract characterisation of the VRE in terms of functionalities to be supported and content to be made available.

The remainder of the paper is structured as follows. Section 2 describes the proposed approach and highlights its adaptivity w.r.t. scenarios that are highly evolving both in terms of user needs and user interface constituents. Section 3 presents an implementation of the approach by discussing the logical architecture of the VRE UI Builder, a software component developed in the context of the D4Science-II EU project. Section 4 describes actual exploitations of the above approach and technology in serving real life scenarios. Finally, Section 5 concludes the paper.

2 Building Virtual Research Environment User Interfaces

In order to properly describe the proposed approach, the assumptions characterizing the application context (cf. Sec. 2.1), the process driving UI building (cf. Sec. 2.2) and the algorithm taking care of constituents arrangement (cf. Sec. 2.3) have to be described.

2.1 Application Context Characteristics

The presentation layer collects all components that expose the VRE functionality to end-users and third-party applications as well. This area represents the *entry point* for using a VRE. In considering the components constituting this area one has to take into account that a comprehensive user interface has to be able to cover all user needs despite their profile and their access device, the functionality, the information object types, formats, and media. Each component must provide a *strong customisation capability* in order to be *easily adapted to the diverse contexts*. VREs have to be accessible by users which may be occasional

and spread worldwide, consequently the best way for end-users to interact with VREs would be through the World Wide Web providing user interface access to all resources belonging to the VRE.

The *resources*, i.e. the services, which partake in a VRE, can be several and change in time. As a result, the VRE presentation layer should face the same variability and dynamicity by adapting to the user interface functionality relative to the services available at a specific time.

A component-oriented approach for building VRE User Interfaces has been adopted. In particular, individual services are delegated the implementation of *User Interface Components* (UICs) that embed the user interface logic of their functionality. The UICs are handled by another service that acts as a "container", conceived as a dynamically configurable service, capable of accessing and combining only those UICs required by the community using the VRE. It is worth noticing that the same service may expose its UIC to different VREs and according to different requirements of the VREs. In that sense, UICs own different states, to be associated to the VREs they are joining. By using a component-oriented approach to compose a web-based user interface we facilitate *reuse*, *sharing* and *customisability*. We do this by (*i*) the automatic combination of generic, reusable web UI components, (*ii*) their distributed deployment provided by the VRE Framework and (*iii*) their context-aware, dynamic invocation, configuration and integration into a homogeneous, web-based container, i.e. a web-portal.

2.2 The User Interface Building Process

According to the assumptions above, every VRE UI consists of a set of UICs that are arranged and hosted by a container. The main issue in this case is given by the fact that the automatic combination of UICs, both graphically and logically, to reflect the chosen functionality set is not trivial. Moreover, the functionalities equipping a VRE are not fixed neither in type nor in form, they result from a VRE specification [1].

It is assumed that for each given functionality there exists a corresponding number of *user interface components* (UICs) that grouped appropriately compose the entire user interface of the functionality. The scenario is depicted in Figure 1. As example, the Search functionality might be used; in this case there could be UICs for performing different type of searches such as full text one in which users just enter a keyword; an advanced one where user may specify logic operators among keywords; a geo-spatial search where latitude and longitude are needed parameters to narrow the search operation, etc.

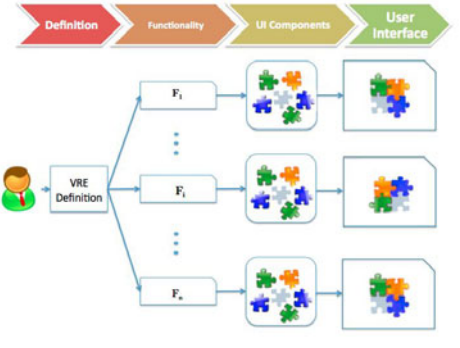


Fig. 1. The VRE UI Construction workflow

In addition to that, it is important to take into account the fact that information objects belonging to a certain VRE content are usually grouped into *Collections*. For this reason a UIC capable of displaying them, usually called *Collections Viewer*, falls into the Search functionality as well. It would not only allow users to perform search operations on specific collections. Rather the same Collection Viewer UIC is used also to perform browsing operations on a given collection, enforcing *sharing* and *reuse* indeed.

It is worth noticing that a single functionality hierarchy level may not be enough: for each *super* functionality coming from the VRE definition there could be other *derived* functionalities all belonging to their parent one. Let's continue with the Search example: in the previous paragraph some UICs that could be combined to provide a user interface for searching into VREs has been listed. The user interface components list was deliberately restricted to UICs capable of defining and submitting search queries, displaying search results must be present as well, it would represent a key-component in this case. This UIC would still belong to the Search functionality though its scope is fairly different from the ones dedicated to query submission. Thus, this UIC belongs to a Search derived functionality, usually denoted as results-presentation while the one dedicate to queries submission may belong to the query-submission derived functionality. The end-user may require more features other than query defining, submitting and results visualisation. For example, in the context of supporting annotations of search results a dedicated UIC which interacts with the annotation service, would belong to the search-annotation derived functionality.

Hence, the previous assumption have to be enriched by adding one hierarchy level in the functionality part. For each functionality there are *derived* ones and these can be arbitrary in number and may vary depending on the contexts. Each macro-functionality exposes its set of derived ones. It is assumed that the derived functionality set is defined a priori within the VRE framework hosting the VRE.

Furthermore, it is assumed that each UIC has to provide the information needed in order to perform logic-graphic combination. Specifically it has to describe its constraints in terms of: (*a*) *size* – each component should supply its minimal width and minimal height (in pixel) to be displayed appropriately; (*b*) *operational context* – each component should supply to which VRE derived functionality it belongs to; a UIC may belong to more than one derived functionality. This information is captured by a *profile*, an XML based specification of the above characteristics.

It is assumed that the VRE interface will be a two-dimensional workspace. The complete VRE User Interface will be composed by a set of rectangular same-size containers, where the smaller rectangular items, i.e. the user interface components, will be packed according to their scope and dimensions.

The set of rectangular same-size containers, i.e. the VRE user interface, will be grouped using tab panels. By knowing each UIC profile, it is possible to get a full knowledge of the UICs set and combine these in largest two-dimensional areas, that will be hosted in one of the available container. In order to combine

UICs efficiently a variation of a Two-Dimensional Bin Packing Problem (2DBP) heuristic algorithm is proposed.

2.3 The FBS-VRE Algorithm

The FBS-VRE algorithm is a variation of the Finite best-strip algorithm [2] adapted for VREs context. The variation mainly regards the fact that in combining UICs it is important to take into account that we aim to compose single portal pages. There are some specific cases not to handle: since each UI Component represents a rectangle in a meaningful part of the final UI, we do not cope with *very tall* or *very flat* rectangles as the lacking of either horizontal or vertical space would not be able to represent a meaningful part of the UI. Moreover, still having in mind the aforementioned goal, we need to *avoid creating crowd user interfaces*. Crowded screens are difficult to understand and, hence, are difficult to use. Experimental results [3] show that the overall density of the screen should not exceed 40%, whereas local density within groupings should not exceed 62%.

Another important feature we need to care of when grouping is to *group things effectively*. Items that are logically connected should be grouped together on the screen to communicate they are connected, whereas items that have nothing to do with each other should be separated. Finally, as the UI Components set for each derived functionality will not be very-large in number, typically in the order of tens, time complexity does not represent an issue. It is worth noting that bin sizes may vary depending on the actual end-user screen resolution, however, as we use pixels as unit of measurement, we assume our bin *width* and *height* as variable parameters for each VRE. Typical values for width vary from 800 to 1280 (pixels) while for the height they can vary from 600 to 1024.

In order not to create crowd interfaces the algorithm applies the UI Components set an horizontal round off. The bin is vertically partitioned in 2, 3, 4 parts. Successively the item, i.e. the UI Component, widths are rounded up to the first partition able to horizontally contain them, i.e. rectangle widths will be updated indeed to 25, 33, 50, 66 or 75 percent of the bin width. To group UI Components effectively instead the algorithm groups them by derived functionality set. In the following, the term *items* is used instead of UI Components for the sake of simplicity.

The FBS-VRE algorithm, for each given VRE derived functionality, starts by sorting the *items* by non-increasing height, and by updating their width rounding up to the first bin vertical partition able to horizontally wrap them. The strip level packing is obtained through the Best-width-fit strategy, looking at all the possible levels and only then pick a best one out of them for the next item to pack. After the first phase the finite solution is obtained by heuristically solving a one-dimensional bin packing problem. Let $h_1, h_2, ..., h_n$ be the heights of the resulting levels, with item sizes h_i and bin capacity H. Through the Best-Fit Decreasing algorithm: initialise bin 1 to pack level 1, and, for increasing $i = 2$ ... n, pack the current level i into the best bin where the remaining horizontal space is minimum, if any; if no bin can accommodate i, initialise a new bin. The pseudo-code of the FBS-VRE algorithm is given in Algorithm 1.

Algorithm 1. FBS-VRE algorithm

```
for each derived-functionality set dfs
    sort the items according to non-increasing h_i values
    for each item i in dfs
        update w_i according to first bin vertical partition capable to wrap i
    comment: first phase
    create new level l_j in the strip
    for each item i in dfs
        if (j > 1)
            choose the level l_best among levels that can accommodate item i according to Best-
fit-width strategy
        if (l_best exists)
            pack item i into l_best
        comment: no levels can accommodate item i
        else
            if item i doesn't fit into l_j
                j := j + 1
            pack the selected item onto the l_j
    end for each
    let λ_h1, λ_h2.....λ_hn be the heights of the resulting levels
    comment: second phase
    determine a finite bin solution by solving the 1BP instance
    having j elements, with associated values λ_h1, λ_h2.....λ_hn and capacity H using best-fit heuristic
end for each
```

3 Implementation of the Proposed Approach

The VRE UI Builder is a core software component for the VRE UI construction that implements the above approach. Given a VRE specification, or VRE definition, it takes care of actually shaping the final VRE user interface. Its output, the UI representation it produces has been designed to be generic so as to avoid the risk of coming up with a tightly coupled software component hardly exploitable in different application scenarios. Its logical architecture is depicted in Figure 2.

A VRE definition [1] is taken as input by the *Specification builder* module whose purpose is the one of providing the *Designer module* the complete specification of the UICs to combine (marked "spec." in Figure 2). The information on and the relations among UICs, needed for defining this specification, are obtained by querying the *User Interface Component Knowledge Base* (UIC KB). Both the Specification builder and the Designer module need to query the UIC KB to get additional information that might be needed in order to come up with a proper VRE UI construction. Specifically, the UIC KB provides two kinds of information: (*a*) the *domain description*, i.e. a description of all the types of UICs available, and (*b*) the *domain context*, i.e. a specification of the operational context for each UIC. The UIC KB component is designed to be generic too. The access to the knowledge contained in the UIC KB is possible through its interface that provides generic methods to retrieve UICs information. In our context the UIC KB has been implemented as a stateful web-service (WS). This WS is part of the VRE framework e-Infrastructure that is in charge of keeping it up-to-date, i.e. maintain the list of the UI component profiles consistent with the availability of resources at VRE creation time. The status of each component

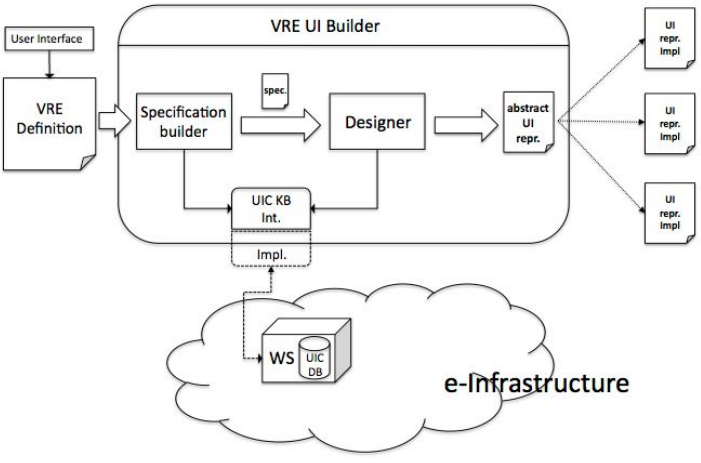


Fig. 2. The VRE UI Builder Logical Architecture

is represented by its XML profile. The list of the UI components to combine provided by the Specification builder module to the Designer is represented in XML form. Its content is made of a hierarchic chain of UI components profiles grouped by super functionality and derived functionality as explained in 2.2.

Successively the Designer module parses the XML spec. obtained from the Specification Builder and uses the FBS-VRE algorithm provided in 2.3 to construct an abstract VRE UI representation. The VRE UI representation provided is deliberately given at an abstract level. It adopts a generic structure that can act as a *carrier* for an arbitrary number of concrete implementations: the developer who wants to bound it to a specific architecture has to realise an architecture specific implementation of the representation.

The implementation of VRE UI representations varies according to the characteristics of the architecture chosen. However, the way a VRE UI representation is defined and its class model structure adopted, makes the the conversion particularly suitable for Portlet container layouts.

4 VRE UI Builder Exploitation into Practice

The VRE UI Builder service has been exploited to serve the needs of communities in different domains ranging from Cultural Heritage to Earth Observation, Biodiversity and High-energy Physics in the context of the D4Science-II project. D4Science-II is one of the main European e-Infrastructure project, involving 11 participants co-funded by the European Commission's FP7 for Research and Technological Development. The project started in October 2009 and has a life of 2 years. D4Science-II continues the path that previous projects have initiated

towards establishing networking, grid-based, and data-centric e-Infrastructures. The goal of the project is to implement an e-Infrastructure for supporting the creation and maintenance of VREs activated on a shared pool of resources ranging from traditional grid resources like computing and storage resources to cloud resources and various types of collections, data sources and application services.

The D4Science-II e-Infrastructure is powered by the gCube System [5], an implementation of a VRE framework concept entirely conceived (*i*) to provide an e-Infrastructure with the management functions needed to properly deal with the potentially huge and heterogeneous set of constituent resources and users, (*ii*) to have on board the minimal set of functions needed to implement VREs giving uniform information management and organisation facilities on a heterogeneous and dynamic information space, and (*iii*) to be open and extensible so as to easily adapt to different application contexts. The resulting framework follows a service-oriented approach and consists of 193 web services, 44 helper software libraries and 76 portlets.

The VRE UI Builder has become an extension of the gCube System. It has been used and validated to support the creation and the management of 8 long-term VREs in a production environment. It has also been used for the creation of short-term VREs: VREs created for given periods of time related to individual events, e.g. a specific VRE was created to perform data analysis in occasion of the recent oil leak occurred in the Mexican Gulf to evaluate the impact on the marine species distribution that live within this geographical area: environmental and biological data have been collected and a specific tool has allowed scientists to compute the probability of marine species distribution within this area. Once work completed, data have been stored in a repository of the e-Infrastructure and the VRE has been dismissed.

These few examples provide an overview of the issues arising during dynamic VRE UI Builder in complex scenarios as those captured by D4Science-II. Despite their complexity, the choices driving the VRE UI Builder design have proven to be effective to successfully satisfy the needs.

5 Conclusion

Virtual Research Environments are internet-based working environments supporting modern research tasks. They are tailored to serve the needs of diverse and evolving user communities. Their development requires innovative approaches including those dedicated to UIs construction. In this paper an approach dedicated to this new research issue has been presented. Such an approach promotes the building of the UIs by dynamically selecting and aggregating the UI constituents needed to serve the specific needs of an application scenario and arranging such components by relying on an heuristic that optimizes the overall UI space consumption. The reference architecture of a software component realizing it in the context of the D4Science-II project has been presented and its implementation briefly discussed. Exemplars of actual VREs developed by relying on this component and conceived to serve diverse needs as those emerging in different

application scenarios ranging from biodiversity to earth observation and cultural heritage have been presented. It is planned to extend and enhance the approach and the related technology by integrating new strategies aiming at improving the overall quality of the UI generation process in terms of *responsiveness*, i.e. the time needed to reflect changes in the VRE UI depending on infrastructure services availability, and *space exploitation*, i.e. the set of constraints and characteristics taken into account while producing the VRE abstract representation.

Acknowledgments. The work reported has been partially supported by the the D4Science-II project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019).

References

1. Assante, M., Candela, L., Castelli, D., Frosini, L., Lelii, L., Manghi, P., Manzi, A., Pagano, P., Simi, M.: An Extensible Virtual Digital Libraries Generator. In: Christensen-Dalsgaard, B., Castelli, D., Jurik, B.A., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 122–134. Springer, Heidelberg (2008)
2. Berkey, J.O., Wang, P.Y.: Two-Dimensional Finite Bin-Packing Algorithms. The Journal of the Operational Research Society 38(5), 423–429 (1987)
3. Bias, R., Mayhew, D.: Cost-justifying usability: an update for an Internet age. Morgan Kaufmann, San Francisco (2005)
4. Blanke, T., Candela, L., Hedges, M., Priddy, M., Simeoni, F.: Deploying general-purpose virtual research environments for humanities research. Philosophical Transactions of the Royal Society A 368, 3813–3828 (2010)
5. Candela, L., Castelli, D., Pagano, P.: gCube: A Service-Oriented Application Framework on the Grid. ERCIM News (72), 48–49 (2008)
6. Carusi, A., Reimer, T.: Virtual Research Environment Collaborative Landscape Study (January 2010)
7. Gietz, P., Aschenbrenner, A., Budenbender, S., Jannidis, F., Kuster, M.W., Ludwig, C., Pempe, W., Vitt, T., Wegstein, W., Zielinski, A.: TextGrid and eHumanities. In: Proceedings of the Second IEEE International Conference on e-Science and Grid Computing, E-SCIENCE 2006, p. 133. IEEE Computer Society, Washington, DC, USA (2006)
8. Hey, T., Tolle, K., Tansley, S.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)

CloudCAP: A Case Study in Capacity Planning Using the Cloud

Joan A. Smith[1], John F. Owen[1], and James R. Gray[2]

[1] Emory University Atlanta, Georgia USA
[2] Kronos, Inc. Chelmsford, Massachusetts USA

Abstract. Emory University Library teamed with a commercial firm to develop a prototype system for using Amazon's EC2 to properly size web application server deployment environments. This approach has been successfully applied to both high-transaction commercial environments with hundreds of thousands of users and to lower transaction digital library environments with hundreds of users. Starting with the same EC2-based product, our goal was to assess whether a similar strategy is practical for an academic library as well as for commercial systems. We examined cloud configuration and deployment costs, test preparation and analysis, and overall feasibility of this approach. Typically, for digital libraries, the user levels are significantly lower, the deployment costs are lower, and the return on investment (ROI) is not as immediately obvious. We conclude that the effort is worth the investment only (a) when there are significant repercussions from under-sizing a newly deployed digital library and (b) sufficient engineering staff are on hand to develop and debug the deployment scenarios.

1 Background

Emory University Library operates over 150 scholarly websites on a dozen small servers ranging from dual-core to quad-core hardware, with average daily visitor (non-robot) numbers in the low to mid-100's according to our Google Analytics reports. Except for rare internet outages and occasional scheduled maintenance, the sites are available and responsive 365x24x7. Using inexpensive, off-the-shelf hardware running a typical "LAMP stack" (Linux, Apache, MySQL and Perl, Python, or PHP), our servers easily handle the low-volume visitor rates typical of academic sites. Our capacity planning has therefore primarily focused on consolidation of services to reduce total hardware investment and administrative overhead for an ever-growing number of sites.

The situation changed dramatically with the launch of the Transatlantic Slave Voyages website[1] (in 2008 and again in 2009), and the new African-Origins website[2] (in 2011). Highly-popular, these digital scholarship sites are actually web applications, more complex than the digital library webs we typically host.

[1] `http://slavevoyages.org`
[2] `http://african-origins.org`

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 110–117, 2011.

News coverage by agencies like the New York Times ramp up site visitor counts by several orders of magnitude: from hundreds per day to tens of thousands per day. Pre-publicity beta deployment proved that our existing approach would lead to embarrassing server failure and site crash. With limited funds, few hardware resources, and a shortage of engineers, we needed to accurately plan for capacity and responsive site performance.

Sophisticated tools to properly size deployment environments have been developed by academic and commercial researchers to ensure satisfactory performance and sufficient capacity of the institutional infrastructure [2],[3]. But using these planning tools is problematic for a small operation like Emory Libraries, in part because we lack the on-site availability of a variety of hardware with which to test candidate configurations. Even where license costs are not prohibitive, effective operation of these tools requires a significant investment in user-training and deployment time, which has so far not been practical for our small engineering team.

A 2009 UC Berkeley report highlighted the Cloud's "elasticity of resources" which allow businesses to meet variability in performance needs without investing in high-cost hardware [1]. However, Cloud-base deployments may be prohibited (e.g., government) or simply unaffordable for 24x7 operations (e.g., Emory University Library). In this case, the Cloud can be used to test potential physical configurations before hardware is purchased and the system deployed.

We spent several years in the commercial sector using the Cloud to size infrastructure for non-Cloud deployments and to identify and eliminate bottlenecks in the application environment. Our strategy, CloudCAP, was based on Amazon's EC2 because of its broad configuration options for computing hardware and operating systems, its ready availability, and Amazon's inexpensive operational fees. From 2008 through 2011 we applied the CloudCAP approach to Emory web applications that had abnormally high visitor rates, i.e., tens of thousands per day instead of just hundreds. This paper discusses our findings.

2 Designing CloudCAP

CloudCAP has a very specific role to play when it comes to capacity planning and performance testing: Provide the most rapid turnaround time possible for evaluating a proposed feature in the environment. Just as improving the edit-compile-debug time improves programmer productivity, reducing the time required to configure, deploy, test and evaluate a system improves product quality through increased tester productivity. Where previously a few basic load tests might be completed, CloudCAP allows more testing to occur, driving the solution towards a configuration that is both economical and responsive. CloudCAP extends the Infrastructure as a Service (IaaS) model to create a Testing as a Service (TaaS) model. It is a web-based application that interfaces with Amazon's storage (S3) and compute Clouds (EC2) to enable testers to rapidly deploy and configure both the application under test (AUT) and the test environment.

2.1 Cloud Definition

The term "Cloud" has been used to describe a wide variety of services ranging from mainframe computing centers to monthly subscription-based services. From our perspective, the Cloud has very specific characteristics. (1) It is sold on demand, not by subscription. (2) It is elastic, i.e., a user can purchase as little or as much as is needed at that time. (3) The hardware is managed by the provider, and (4) it is characterized by rapid provisioning and deployment of systems, near real-time for some Cloud services. These features are what make the Cloud attractive as a platform for capacity planning and performance testing.

2.2 CloudCAP Architecture

CloudCAP has three responsibilities. First, it facilitates rapid application deployment of both the AUT and the test agents. All software components required for both AUT and testing are preinstalled in the operating systems images we create for use by the solution. The image contains preinstalled copies of all software required by any of the nodes, such as database engine, application server, web server, etc, allowing us to maintain a single Amazon Machine Image (AMI) per operation system, independent of the type of node the image will become. All node types launch from the same image. The second responsibility is configuration of the nodes in the cluster. This configuration cannot happen until the nodes have been booted and are connected to the network. To avoid excessive polling, this configuration control is inverted, with each node asking CloudCAP for configuration data at two distinct stages in its boot process. The node's configuration scripts are hooked into the *rc.local* script on Unix style operating systems, and in the startup group policy on Windows operating systems. Finally, the tool is responsible for monitoring and aggregating performance data from the cluster.

Figure 2 gives an overview of the process from the perspective of a sequence diagram. The entire sequence takes only minutes to complete, whereas deploying physical hardware with a new configuration takes us several hours or longer. Stage-1 configuration has the nodes requesting from CloudCAP the node's specific configuration data. Once all nodes report to CloudCAP that Stage-1 configuration is complete, Stage-2 begins. Nodes learn the IP addresses of their peers enabling an application server to initialize a connection pool back to the database.

Consider a two machine cluster with a MySQL database server and a Java Servlet container, both on Windows 2003 Server. CloudCAP would first instruct EC2 to launch two copies of CloudCAP's custom Windows 2003 Server image. Once the first machine boots, its Startup Group Policy instructs it to connect to CloudCAP, and retrieve its Stage-1 instructions. The node starts configuring MySQL, and loading any customer specific data it needs. Similarly, the second node can start configuring Apache Tomcat, but it cannot actually launch it until Stage-2, when it has the address of the database connection. Once finished with Stage-1 processing, both nodes begin polling CloudCAP for peer data.

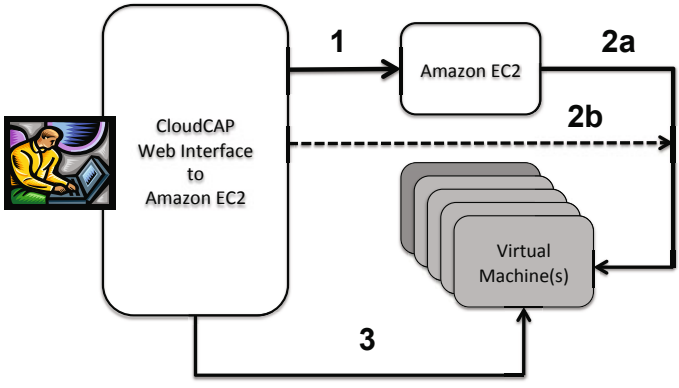


Fig. 1. CloudCAP is a front-end for Amazon's EC-2 service. A single tester can configure & instantiate a variety of tests from a single CloudCAP instance. Step (1) User initiates CloudCAP via the web interface. Step (2a) EC2 launches the virtual machines configured by the user's selection of options (2b). In Step (3) the user monitors the machines and gathers data from the performance tests.

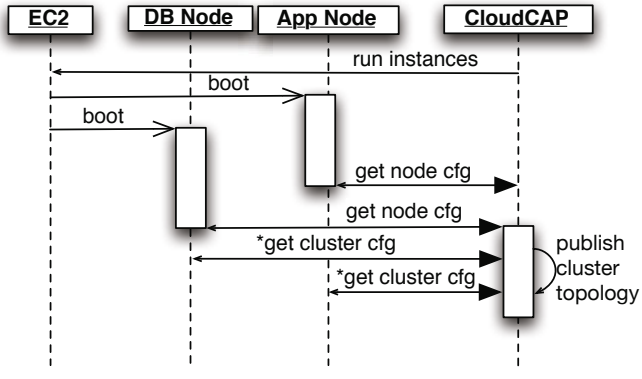


Fig. 2. Abbreviated sequence diagram of CloudCAP instantiation & operation

Once peer data is provided, cluster configuration concludes with the second node configuring the database connection pool and starting Apache Tomcat.

2.3 Using CloudCAP

CloudCAP can be used for both performance and functional testing. It also serves to prove the deployability of the packaged software: If it does not install properly or components are missing, this fact is immediately discovered during CloudCAP initialization. See Figure 3 for a screen shot of the simple initialization interface. The effort of writing the test scripts is the same in both traditional and cloud-based testing. However, with CloudCAP an additional up front effort

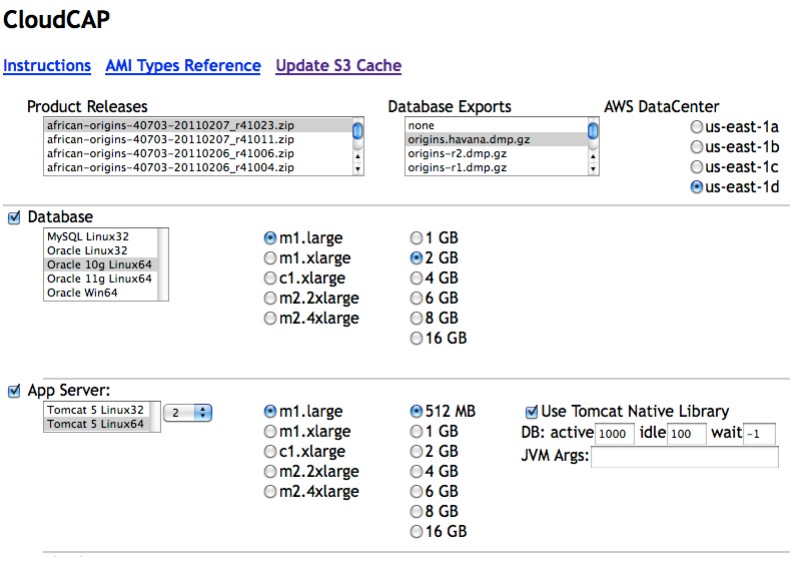


Fig. 3. Part of the CloudCAP web interface to configuration options. Product release list comes from the software's version control repository.

is required to script the deployment of the AUT into the CloudCAP framework. The primary benefit of CloudCAP is that once the deployment has been scripted, the full range of tests can be conducted by a single test engineer, and the updates and retests can be accomplished in a matter of minutes or hours instead of days or weeks. For the Library, with a project queue far longer than its staff capacity, thorough product testing is highly resource constrained. Off-loading the test setup, operation, and results reporting to a one-button system like CloudCAP has helped us prepare more complex sites for successful launch.

2.4 Case Study

The Africans-Origins website launch was planned to include major news media coverage, meaning the site would experience abnormally high traffic volume, estimated at 12,000 to 20,000 users per day based on our previous experience with the Transatlantic Slave Voyages website. We use page response time and page throughput as key metrics to measure performance since this has proven a reliable indicator of our corporate and government customers' satisfaction. The page response times are measured in seconds per page, and throughput in pages per second. For African-Origins, our target performance level was a page response of no more than 0.5 seconds per page and a throughput of ≥ 20 pages per second at peak unique visitor levels. We used CloudCAP both to estimate the capacity of our best in-house server and to identify the minimum configuration that would meet our performance goals.

The Library is not equipped with the infrastructure to perform comparative hardware stress-tests, so we relied on CloudCAP to identify performance bottlenecks and recommend server configurations. The key to the evaluation was CloudCAP's ability to quickly configure machine images and clusters with the target environment, and the ability to rapidly deploy, and re-deploy, the African-Origins site into the test environment so as to support quick assessment of environment and application changes. Using CloudCAP, machine images and a single cluster were configured to support the application environment and the test agent environment. Test agent scripts were created that simulated unique visitor access, and emphasis was placed on the sections of the application that were expected to be areas of high-traffic. Initial testing was conducted to evaluate the proper configuration of the test environment and the deployment of the African-Origins application within the environment. After verifying that CloudCAP was properly configuring the test environment, an initial performance test was run for the purpose of establishing a baseline set of metrics.

The initial test run revealed that the application server CPU utilization was nearly 100% and that system saturation occurred at a level of only 30 unique visitors, far below acceptable limits. Further tests indicated that the Levenshtein search process was the primary bottleneck. A refactoring of the search implementation to improve its efficiency increased site performance to acceptable levels. Additional tests revealed a potential saturation point with respect to database access and ongoing refinements to the application were made to minimize its impact on system performance. We continued to conduct "what-if" analysis on the site, adjusting allocated resources (CPU, RAM, etc.) until we reached an optimal configuration. Each "what-if" session took only minutes of the tester's time, whereas performing those same hardware adjustments and redeployments on systems in our labs would have taken many hours per configuration change.

3 Results

We found a number of positive results arising from CloudCAP. It allows for what-if scenarios in both topology and in tuning parameters including specific details like RAM allocations, and number of CPUs. Options can be configured to meet domain-specific needs, allowing for more flexible scenario testing. CloudCAP supports not only a LAMP stack but also Java Enterprise stacks (Tomcat, Oracle, Apache), and Windows environments (Server 2008, IIS, SQL Server). Cloud-based capacity planning and assessment provides a cost-effective means to:

- Quickly configure varying combinations of hardware and operating systems that simulate the production environment
- Quickly deploy load-test agents that can simulate varying levels and types of customer usage patterns
- Quickly conduct "what-if" analysis to determine if variances in system configuration can improve overall application performance.

Once the test environment has been configured, robust testing of the product can be conducted with minimal support from QA staff, resulting in a faster turn-around time between testers and developers and culminating in an improved end-product. The on-demand nature of the Cloud is a perfect match for the bursty nature of load testing and for overall capacity planning whether the eventual deployment will be local or in the Cloud itself. On the other hand, using CloudCAP requires a substantial engineering investment by the institution. In part this is because the API for integrating tests into the CloudCAP framework is still immature, and configuring CloudCAP to test a new product may outweigh relative benefits. In a commercial environment, the benefit accrues over a span of several product releases, encompassing the entire lifecycle of the product. In the library environment, a product may have only a single release. New sites often see a very heavy load during the initial deployment which then subsides into the more typical low-level page hits, so the institution might find it cheaper to simply "throw hardware at the problem." After the initial surge completes, engineering focus shifts to consolidation onto a shared service environment, where many applications co-exist on common hardware. CloudCAP can be used to validate that the hardware configuration of the shared environment has sufficient capacity to support consolidation.

One challenge is mapping the EC2 hardware to equipment that can be purchased from commercial vendors. Amazon characterizes compute power in terms of a synthetic benchmark, *compute units*. Amazon defines one EC2 Compute Unit as providing the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.[3] However, because there is no standard benchmark to evaluate more modern equipment in terms of compute units, translating an EC2 environment into practical hardware is largely guess work.

The next difficulty centers around the variability in hardware provided by Amazon. While in theory a test might be running on a 12-compute-unit box, other virtual machine tenants on the same physical hardware might coincidentally also be compute-intensive. Although Amazon attempts to virtualize equally, they have no predefined knowledge of the usage scenarios of the many tenants they colocate on the same physical hardware. In our experience, multiple test runs are necessary to even-out the random effects of colocation with other tenants.

4 Future Work

The Cloud-based testing approach has proven invaluable for integrating load testing into the workflow, but its use has highlighted some shortcomings. A high initial investment is required for each new application introduced into the CloudCAP environment. The application plugs into the CloudCAP service through a still evolving API. This API needs to be matured so that it provides a flexible and stable interface by which an application can specify its full installation process and implement its own configuration into the clustered environment. Even

[3] Cf. Amazon's description at http://aws.amazon.com/ec2/instance-types/

though some standardization of application installations exist such as the Java WAR and EAR specifications, many of the details of deploying applications in a clustered environment remain painfully manual.

Performance monitoring is an important aspect of any load testing effort. In the CloudCAP environment it is still very ad-hoc, and is based on Amazon's Cloud Watch monitoring tool and a collection of operating system specific tools (Windows Performance Monitor, top, ps etc). An ideal system would aggregate tools from the Cloud provider together with operating system tools running on the individual nodes similar to the Ganglia tool.[4] Such an aggregation would allow rapid assessment of performance, further reducing the total testing time of a particular permutation.

A considerable amount of work still needs to be done before CloudCAP can be released as an Open Source product. In part the problem is one of practicality, since each institution's environment needs to be customized within the CloudCAP framework. In other words, there is not as much general reusability as we had hoped. Nonetheless, the Cloud-based testing approach shows promise for organizations with strong in-house engineering expertise and the need to support a wide range of performance requirements or deployment environments.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: Above the clouds: A Berkeley view of cloud computing. Tech. Rep. UCB/EECS-2009-28, University of California at Berkeley (February 2009)
2. Bagchi, S., Hung, E., Iyengar, A., Vogl, N., Wadia, N.: Capacity planning tools for web and grid environments. In: Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools. ACM, New York (2006)
3. Smit, M., Nisbet, A., Stroulia, E., Edgar, A., Iszlai, G., Litoiu, M.: Capacity planning for service-oriented architectures. In: Proceedings of the 2008 Conference of the Center for Advanced Studies on Collaborative Research: Meeting of Minds, pp. 11:144–11:156. ACM, New York (2008)

[4] `http://ganglia.info`

Query Operators Shown Beneficial for Improving Search Results

Gilles Hubert[1], Guillaume Cabanac[1],
Christian Sallaberry[2], and Damien Palacio[2]

[1] Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
[2] Université de Pau et des Pays de l'Adour, LIUPPA ÉA 3000
Avenue de l'Université, BP 1155, F-64013 Pau cedex

Abstract. Search engines allow users to retrieve documents with respect to a given query. These provide advanced search options, such as query operators (e.g., `+term`, `term^10`). Previous work studied how query operators are employed by end-users. In this paper, we study the extent to which using query operators may lead to improved results, regardless of specific users. We hypothesize that the proper use of query operators improves search results. To validate this hypothesis, we present a methodology relying on standard IR test collections. We applied this methodology to TREC-7 and TREC-8 test collections with five IR models implemented in the Terrier search engine. Experiments show that queries enriched with operators give an improvement in effectiveness up to 35.1% over regular queries. This result suggests that end-users would benefit from using operators more often.

Keywords: Information Retrieval, Search Engine, Query Operators, Effectiveness.

1 Introduction

Digital Libraries (DL), such as Europeana [15], aim to collect and give access to huge amounts of multimedia documents. People may either browse these repositories or retrieve documents matching their interests, thanks to a search engine. In the latter case, users have to translate their information need into a sequence of terms, called a query. For instance, a scientist looking for research projects funded in the DL domain may issue the following query: [`digital library research project funding`].

For reducing the mismatch between the user's cognitive model of the need and the produced query, search engines offer query operators [17], as a way to specify the role of (group of) terms. These comprise boolean operators (e.g., `AND`, `OR`, `NOT`), expressions surrounded by quotation marks (e.g., `"digital libraries"`), proximity operators (e.g., `NEAR`), boosting operators, and so on. In the previous example, the scientist may expect better results when wording his/her query as: [`"digital library" +research project funding^4`]. This refined query

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 118–129, 2011.

better conveys the intent of the user, since the search engine is instructed that the research aspect is mandatory, and the funding aspect should be considered as a strong evidence for ranking documents in the result list.

Search engines foster the use of query operators to improve retrieval quality, as reported in [4]. The real use of query operators was studied in [17,18,4,23,1] by analyzing logs of popular search engines, such as Altavista, Excite, Google, MSN Search, and Yahoo!. Researchers found that queries with operators constitute up to 20% of all submitted queries. This recognized use raised questions about the effects of operators on search accuracy. While operators are expected to improve search engine effectiveness, as underlined by White and Morris [23], experiments by Eastman and Jansen [4] showed that the outcome may be 'relatively small.' The reasons behind this poor improvement may be related to various factors ranging from users to Information Retrieval (IR) models implemented in search engines. Although previous work was devoted to understanding the 'user' factor [17,18,11,4,5,1], we failed to find any research investigating the 'system' parameter. We therefore analyze in this paper the potential of effectiveness improvement yielded by operators, regardless of specific users. We address hypothesis $\mathcal{H}$: *the proper use of query operators improves search results.*

The paper is organized as follows. In Sect. 2, we review the literature devoted to query operators as featured by search engines. We stress that studies to date did not measure the potential of effectiveness improvement when properly using query operators. In Sect. 3, we present the proposed evaluation methodology involving experiments with standard IR test collections. In Sect. 4, we report the results of the experiments that we conducted with the TREC-7 [20] and TREC-8 [21] test collections. These validate $\mathcal{H}$: *the proper use of query operators does improve search results.* We conclude the paper in Sect. 5, and give insights into future work.

2 Related Work: Operators in Search Queries

Studies of operators in queries submitted to search engines fall into the two categories discussed in the following sections.

2.1 Usage of Query Operators

Several studies reported the use of query operators for common search engines. Analyzing various query logs with different characteristics (number of queries, users, crawling timespan), researchers found the following proportions of queries with operators — due to space limitation, we emphasize on most recent research:

- For Altavista, Silverstein et al. [17] found 20.4% in 1999.
- For Excite, Jansen et al. [10] found 24.1% in 2000, while Spink et al. [18] found 14.5% in 2001.
- For Google, MSN Search, and Yahoo! altogether, White and Morris [23] found 1.12% in 2007. Notice, however, that this study was limited to four operators (i.e., `+`, `-`, `"..."`, and `site:`).

In addition to these quantitative studies, other research was concerned with qualitative analysis related to users. Hölscher and Strube [8], as well as Lucas and Topi [11], found that expert users recourse to query operators more frequently than the average user. Using query operators is a trait that one would expect from expert searchers, according to White and Morris [23]. Jansen et al. [10] point out that average users "are certainly not comfortable with Boolean operators and other advanced means of searching."

When present in queries, operators are used in a "semantically appropriate manner," according to Eastman and Jansen [5]. Users tend to use more operators when facing complex information needs or having difficulty in finding information [1]. Overall, query operators were found to be used more in dedicated search engines (e.g., online DL catalogues) than in web search engines [9].

2.2 Benefits Brought by Using Query Operators

Beyond measuring the proportion of queries with operators in search engine query logs, a few studies investigated their effects on retrieval effectiveness. Eastman and Jansen [4] stated in 2003 that only few studies compared retrieval results using query variants (i.e., with and without operators). Since then, White and Morris confirmed that observation: query operators "have generally been overlooked by the research community in attempts to improve the quality of search results" [23].

One prominent study measuring the effect of query operators on result accuracy was conducted by Eastman and Jansen [4]. Their experiment involved two sets of queries (A and B). Set A contained 100 original queries with operators (`AND`, `OR`, `MUST APPEAR`, and `PHRASE`), which were extracted from Excite logs. Set B contained the 100 original queries with all operators removed. Finally, queries from sets A and B were submitted to three search engines: AOL, Google, and MSN Search. The top 10 documents retrieved by each search engine, for each query, were judged by 4 experts on a 4-point scale. Documents marked with an average score of 3 or higher were considered as relevant to the query. Among other measures, averaged 'relative precision' P@10 (i.e., number of relevant documents in the top 10) was calculated for sets A and B. The researchers showed that the recourse to query operators (set A) did not yield statistically significant improvement over operator-free queries (set B). They concluded that "the use of most operators had no significant effect on ... relative precision."

Eastman and Jansen's [4] conclusions may lead searchers to get rid of query operators. Nevertheless, we wonder why this study focused on queries with operators in the first place, since these only represent up to 20% of all submitted queries. In addition, these are known to be more complex than average queries [1]. That is the reason why we ask ourselves, in this paper, whether their conclusions still hold for the 80% remaining queries that users formulate without operators. To do that, we intend to evaluate the improvement in effectiveness yielded by refining regular queries with operators. The next section introduces the methodology that we designed for that purpose.

3 Methodology: Assessing the Effects of Query Operators

We designed an evaluation methodology to test $\mathcal{H}$, that is: *the proper use of query operators improves search results*. Two research questions arose when trying to validate this hypothesis:

- Q_1. What is the maximum gain in effectiveness that one can expect by enriching a query with operators only (no term modification or addition)?
- Q_2. Do users succeed in formulating queries with operators, so that these lead to a significant gain in effectiveness?

If we notice no possible gain when using query operators (Q_1), we obviously cannot expect users to get better results when having recourse to them (Q_2). Hence, the answer to Q_2 depends on the answer to Q_1. We thus focus on answering Q_1 in this paper. The proposed methodology is illustrated in Fig. 1. It involves the four stages that we detail in the following sections.

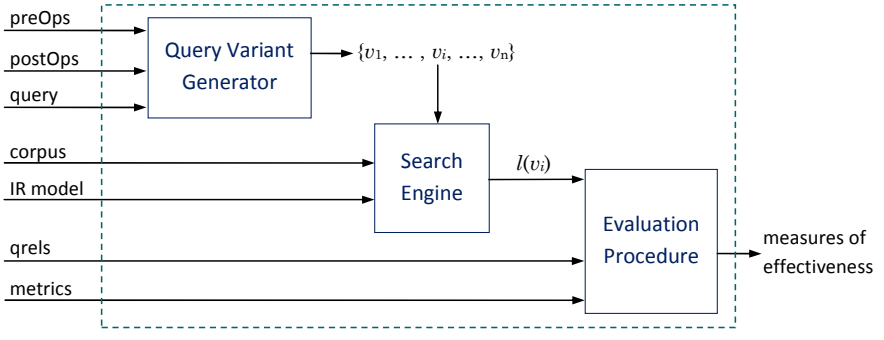


Fig. 1. Illustration of the methodology for assessing the effects of query operators

3.1 Selection of an IR Test Collection

An IR Test Collection allows researchers to run experiments for evaluating their search engines [16]. It is comprised of four components:

1. The *corpus* is a set of documents to be indexed by the search engine.
2. A set of n *topics* represents user information needs. Each topic may be worded as:
   - A *title*: a sequence of two or three terms (in general) that would be submitted as a query to a search engine by an average user.
   - A *description*: a few sentences describing the user's information needs in plain text.
   - A *narrative*: a longer text than the description, which gives insights into the intent of the user, and unambiguously states what information is relevant or irrelevant for the searcher.

3. The *qrels* a.k.a. query relevance judgments, or gold standard. Usually, experts identify and mark documents according to their relevance to each topic. These marks/ratings may be binary (i.e., nonrelevant vs. relevant) or gradual (e.g., ranging from nonrelevant to relevant on a 5-point scale).
4. The *metrics* allow the measurement of the accuracy of the results retrieved by a search engine. As detailed in [2], common metrics are Precision, Recall, F1, Average Precision, and so on. Taking as input (i) the given topic and (ii) the result list produced by the search engine, a metric gives a numerical value representing the effectiveness of the search engine in retrieving relevant documents to the user.

Several IR Test Collections were produced by IR initiatives and then released to the community for research purposes. As the prominent evaluation initiative in IR, TREC [22] has been providing many collections [7] since 1992. Note that other initiatives exist with similar purposes (e.g., CLEF, NTCIR).

3.2 Generation of Query Variants with Tested Operators

We intend to check that any given user's query can be rewritten with operators, such that it leads to more accurate search results. In the remainder of the paper, we call 'preOps' the operators prefixing a query term (e.g., ∅, `+`), and 'postOps' the operators postfixing a query term (e.g., ∅, `^2`, `^10`). For each *topic* in the selected Test Collection, we consider the associated *title* (or *description*, or *narrative*) comprising t terms. Then, a Query Variant Generator is used to generate $\omega = (b \cdot a)^t$ variants with the b given preOps, and the a postOps to be tested. These are denoted $\{v_1, \ldots, v_i, \ldots, v_n\}$ in Fig. 1.

Example 1. Let us consider a *topic* from a Test Collection, with the following *title*: [`1974 Turing award recipient`]. Generating variants for the $b = 2$ preOps $\in \{\varnothing, \texttt{+}\}$, and $a = 3$ postOps $\in \{\varnothing, \texttt{\^{}2}, \texttt{\^{}10}\}$ leads to $\omega = (2 \cdot 3)^4 = 1{,}296$ query variants illustrated in Table 1.

Table 1. Excerpt of the 1,296 query variants generated with prefix operators $\{\varnothing, \texttt{+}\}$ and postfix operators $\{\varnothing, \texttt{\^{}2}, \texttt{\^{}10}\}$

Variant #	Query variants generated with preOps and postOps			
1	`+1974`	`Turing`	`award`	`recipient`
2	`1974`	`+Turing`	`award`	`recipient`
⋮	⋮	⋮	⋮	⋮
1,295	`+1974^10`	`+Turing^10`	`+award^10`	`+recipient^2`
1,296	`+1974^10`	`+Turing^10`	`+award^10`	`+recipient^10`

3.3 Retrieval with Initial Query and Generated Variants

During this third stage, the documents from the *corpus* are indexed by a Search Engine. Then, it is run according to a given IR model, which governs the way queries are matched with documents. TF·IDF and OkapiBM25 may be cited as examples of prominent IR models. Since detailing how these models operate is beyond the scope of this paper, we refer the reader to [3, chap. 7] for a comprehensive presentation of this topic.

Finally, the initial (operator-free) query q, and the generated ω query variants v_i (with operators) are submitted to the search engine. In the remainder of the paper, $l(q)$ denotes the list of documents retrieved for query q, and $l(v_i)$ denote the list of documents retrieved for a query variant v_i. Note that all lists of documents are ranked by decreasing RSV (i.e., Retrieval Status Value: the relevance score estimated with respect to the query).

3.4 Data Analysis: Measuring Effectiveness Variations

The fourth stage of the methodology is concerned with the analysis of search results. An Evaluation Procedure applies a metric m to the document list $l(q)$ or $l(v_i)$, and to the *qrels* $j(q)$ associated with the tested initial query q. The value of this metric $m(l(q), j(q)) \in [0, 1]$, also called 'measure,' represents the extent to which query q yielded relevant results. Similarly, $m(l(v_i), j(q)) \in [0, 1]$ represents the extent to which query variant v_i yielded relevant results. According to these measures, one may report per topic analyses, as well as global analyses.

Per Topic Analysis. For a given evaluation metric m (e.g., Recall, Average Precision), a given initial query q, and its variants $v_{i \in [\![1,\omega]\!]}$, one gets $\omega + 1$ measures. These data values represent the outcome when applying query operators to initial query q. Among them, the *maximum value* $\max_{i=1}^{\omega} m(l(v_i), j(q))$ is the best performance reachable by using operators properly.

In addition, one may study the distribution of effectiveness data values thanks to the 'boxplot' visualization [19,24]. As shown in Fig. 2, a boxplot (a.k.a. box-and-whisker diagram) summarizes several descriptive statistics. The interquartile range (IQR) spans the lower quartile to the upper quartile. The middle 50% of the ranked data lies in the IQR. It is represented as a *box* (central rectangle), which shows the spread of data values. The median is shown as a *segment* inside the box. This is the middle half of the data values, and allows one to assess the symmetry of the distribution. The *whiskers* extend from the ends of the box to the most distant value lying within $1.5 \times$ IQR. Larger and lower values are considered as outliers; these are plotted with *black circles*.

For a given *topic*, we may finally compare the effectiveness of the initial query q with the effectiveness of the best query variant v_i. Equation (1) computes the percent gain yielded by query operators.

$$g(q, v_i) = 100 \times \left(\frac{m(l(v_i), j(q))}{m(l(q), j(q))} - 1 \right) \tag{1}$$

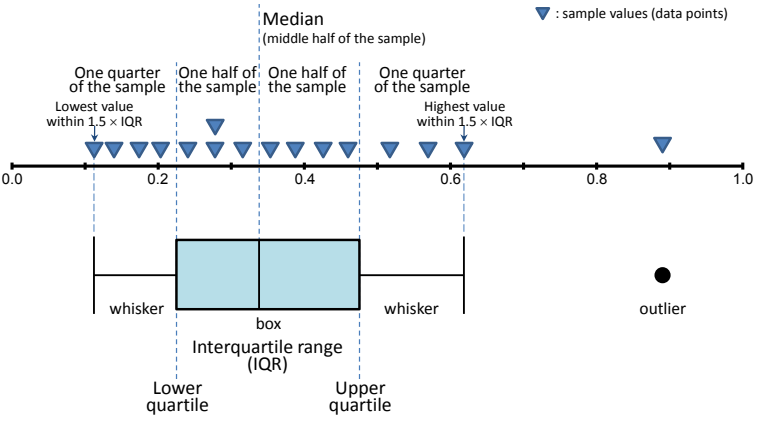


Fig. 2. Example of data values and associated boxplot

Global Analysis. Controlling for 'topic effect' is usually done by evaluating the search engine with n topics, and then averaging out the individual n effectiveness scores (e.g., AP meaning Average Precision) into a global score (e.g., MAP meaning Mean Average Precision). This practice was shown to give stable conclusions with a least $n = 25$ topics, while $n = 50$ is the standard at TREC [2].

The comparison between MAP of initial query q, and MAP of best query variant v_i allows the checking of $\mathcal{H}$. For validating this hypothesis, we must show that (i) v_i is more effective than q, and that (ii) the observed difference is statistically significant with regard to Student's two-tailed paired t-test. According to the resulting p-value, the two data samples are said to be statistically different when $p < 0.05$. The interested reader is referred to [16] for in-depth coverage of statistical testing in IR.

4 Experiments and Results

We applied the devised methodology to TREC-7 [20] and TREC-8 [21] standard test collections. They provide a corpus of newspaper articles; this corresponds to the kind of documents that a DL would index (contrary to other TREC test collections providing web documents). Moreover, they provide $n = 100$ topics covering various subjects, allowing us to conduct significance testing.

We tested two query operators: (i) the 'must appear' prefix operator (`+`), as it is considered as popular [8] and "easy to employ and available" [4], as well as (ii) the boosting postfix operator (`^N`), as we found no study to date on this operator. Among search engines supporting these both features (e.g., Lemur [12], Lucene [6], Terrier [13]), we used Terrier version 3.0, which provides several IR models. In order to check the 'model effect' on results, we conducted experiments with the following five models: BM25, DFR_BM25, InL2, PL2, and TF_IDF. We queried Terrier with the *title* part (see Sect. 3.1) of the 100 topics, the 9,953 variants generated for TREC-7, and the 11,203 variants generated for TREC-8.

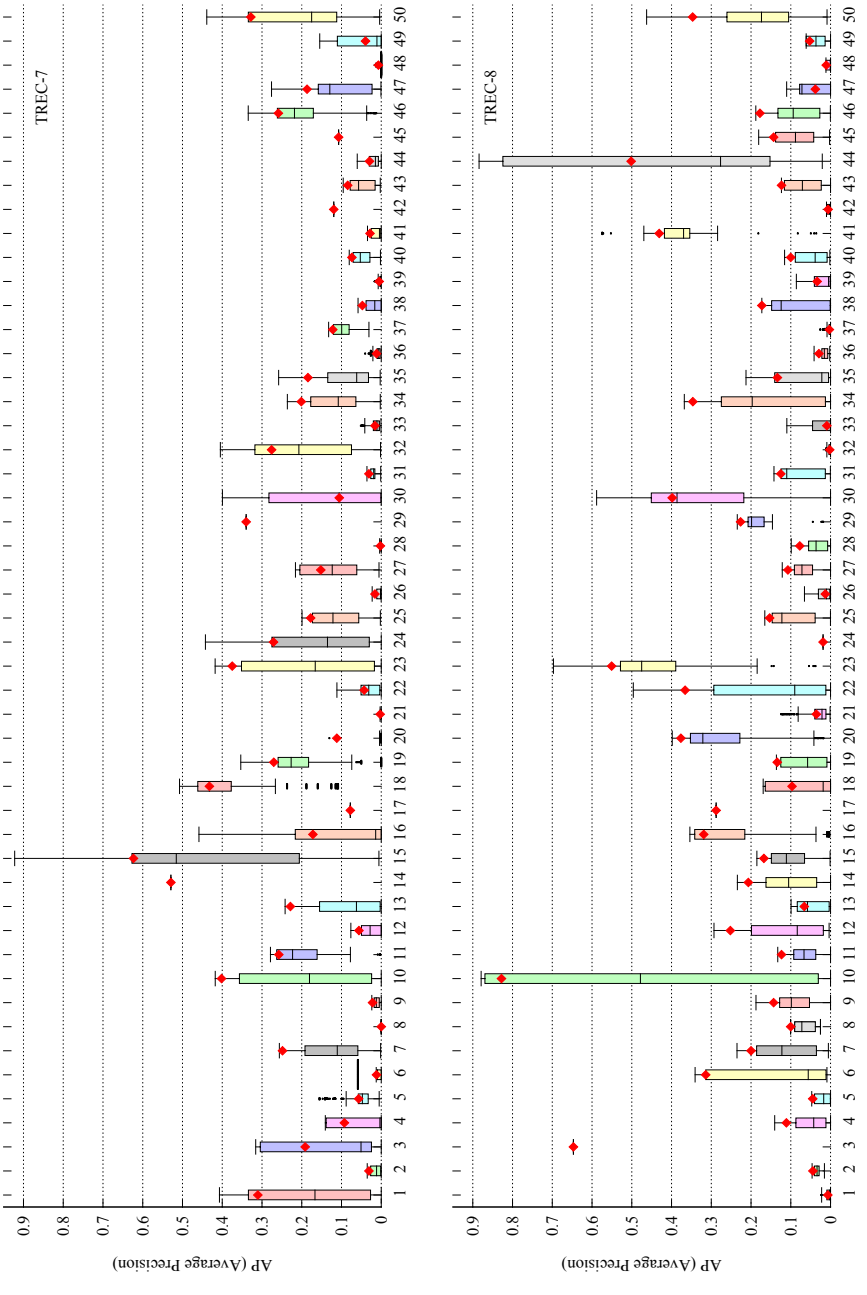


Fig. 3. Average Precision (AP) boxplots showing potential for improvement with query operators on 100 queries (TREC-7 and TREC-8) with model PL2 of Terrier [13]. Diamonds ◆ show the AP of TREC's initial operator-free query.

4.1 Per Topic Analysis

We report in Fig. 3 the per topic effectiveness (AP) of the baseline (initial query q without operators, plotted as red diamonds ◆), as well as the boxplots summarizing the APs of all query variants. We used the default model provided in Terrier [13], namely PL2. Overall, regarding both TREC-7 and TREC-8 results, the baseline AP is highly variable, suggesting that these test collections feature a mixture of easy and hard topics.

Regarding mono-term queries (e.g., Topic 14 for TREC-7 or Topic 3 for TREC-8), we noticed no difference between AP of q and AP of v_i. This result was expected since a mono-term is mandatory *per se*, and boosting has effect with multi-term queries only.

For the remaining queries, the baseline AP generally lies over the median (i.e., segment inside the central rectangle) in Fig. 3. This suggests that most variants led to worse AP than the initial query q did. Nevertheless, there is always a query variant v_i whose AP equals or outperforms the AP of the baseline query q. This observation tends to support the hypothesis $\mathcal{H}$. This improvement does not seem to depend on the AP of the baseline: there is way for improvement for poor baselines, as well as for strong baselines.

4.2 Global Analysis

In addition to per topic analysis, we quantify the observed performance differences thanks to a global analysis covering the n topics of TREC-7 and TREC-8. We measured the MAP of the baseline (queries without operators), and the MAP of the best variant per topic (queries with operators), for each of the five tested IR models. Then, we computed the gain brought by operators (Δ in percent, as reported in Tables 2-3) to validate hypothesis $\mathcal{H}$.

We conducted three experiments: (i) with operators $\{\varnothing, +\}$, (ii) with operators $\{\varnothing, \hat{}10, \hat{}20, \hat{}30, \hat{}40, \hat{}50\}$ featuring arbitrarily chosen weights, and (iii) a combination of the two.

Table 2. Evaluation results with the 'must appear' operator (+)

Model	TREC-7			TREC-8		
	MAP			MAP		
	Baseline	VOP	Δ(%)	Baseline	VOP	Δ(%)
BM25	0.1677	0.1836	9.5**	0.1957	0.2154	10.2*
DFR_BM25	0.1683	0.1843	9.5**	0.1965	0.2162	10.0*
InL2	0.1710	0.1852	8.3**	0.1996	0.2172	8.8*
PL2	0.1554	0.1826	17.5**	0.1840	0.2106	14.5**
TF_IDF	0.1674	0.1833	9.5**	0.1964	0.2158	9.9**

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

In Table 2, query variants with the plus operator (VOP) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. In Table 3, queries with the boosting operator (VOB) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. Overall, the boosting operator yields better results than the 'must appear' operator.

Table 3. Evaluation results with the boosting operator (`^N`)

Model	TREC-7			TREC-8		
	MAP			MAP		
	Baseline	VOB	$\Delta(\%)$	Baseline	VOB	$\Delta(\%)$
BM25	0.1677	0.2027	20.9**	0.1957	0.2312	18.1**
DFR_BM25	0.1683	0.2034	20.9**	0.1965	0.2316	17.9**
InL2	0.1710	0.2059	20.4**	0.1996	0.2352	17.8**
PL2	0.1554	0.1926	23.9**	0.1840	0.2173	18.1**
TF_IDF	0.1674	0.2026	21.0**	0.1964	0.2312	17.7**

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

In Table 4, queries with operator (VOPB) always overcome the baseline, whatever the IR model. Notice that the differences are statistically significant. Overall, the combination of 'must appear' and boosting operators yields best results, up to an improvement of 35.1%. This is a material improvement, which would be even larger if we had got rid of mono-term queries, since they do not benefit from any operator.

Table 4. Evaluation results with boost and 'must appear' operators (`+` and `^N`)

Model	TREC-7			TREC-8		
	MAP			MAP		
	Baseline	VOPB	$\Delta(\%)$	Baseline	VOPB	$\Delta(\%)$
BM25	0.1677	0.2132	27.1**	0.1957	0.2381	21.7**
DFR_BM25	0.1683	0.2133	26.7**	0.1965	0.2387	21.5**
InL2	0.1710	0.2144	25.4**	0.1996	0.2407	20.6**
PL2	0.1554	0.2099	35.1**	0.1840	0.2288	24.3**
TF_IDF	0.1674	0.2131	27.3**	0.1964	0.2383	21.3**

Statistical significance is denoted by '*' for $p < 0.05$ ('**' for $p < 0.01$)

4.3 Discussion of Results

In our experiments, Terrier [13] was considered as a black box. To our knowledge, most IR models do not specify how to handle mandatory terms, as well as

boosted terms. Terrier, however, implements this feature, which suggests that it performs a specific computation.

Regarding the reported results for the boosting operator (Tables 3–4), we selected boosting weights arbitrarily. Other values may have given different results. We leave to future work the study of boosting weights on effectiveness.

5 Conclusion and Future Work

Previous work considered the use of query operators in common search engines. Eastman and Jansen [4] notably studied whether queries with operators yield similar effectiveness with respect to counterparts without operators. They reported a limited improvement due to operators, which questions the return on investment that users may grant to query operators. We wondered if this poor effect was due to users or search engines.

In this paper, we considered the majority (80%) of queries submitted to search engines: those without operators. We stated hypothesis $\mathcal{H}$: *the proper use of query operators improves search results*. We designed a methodology to validate $\mathcal{H}$ through the use of standard IR test collections, and the generation of query variants with operators. We applied this methodology using TREC-7 and TREC-8 test collections. Experiments showed that TREC's initial query can always be improved by refining it with 'must appear' and boosting operators. The observed gain — up to 35.1% — is statistically significant, whatever the tested IR model and collection. This suggests that, when properly used, users benefit from refining queries with such operators. Indeed, query operators convey information instructing the search engine about requirements and preferences (as expressed in the *narrative* part of topics) that would remain implicit otherwise.

Directions for future work include, in the short term, experimenting our methodology in various contexts (e.g., additional IR collections, IR models, query operators). In the medium term, we plan to address Q_2 stated in Sect. 3: do users succeed in formulating queries with operators, so that these lead to a significant gain in effectiveness? In addition, we should study other factors involved when retrieving information [11], such as the number of terms used, and the selection of terms. In the long term, we may study operator use and effects for retrieval involving more than the topical dimension of information. This is notably the case of geographic IR [14] involving spatial and temporal dimensions in addition to the topical dimension of information.

References

1. Aula, A., Khan, R.M., Guan, Z.: How does search behavior change as search becomes more difficult? In: CHI 2010: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 35–44. ACM, New York (2010)
2. Buckley, C., Voorhees, E.M.: Retrieval System Evaluation. In: Voorhees and Harman [22], ch. 3, pp. 53–75

3. Croft, W.B., Metzler, D., Strohman, T.: Search Engines: Information Retrieval in Practice. Addison-Wesley, Reading (2010)
4. Eastman, C.M., Jansen, B.J.: Coverage, relevance, and ranking: The impact of query operators on web search engine results. ACM Trans. Inf. Syst. 21(4), 383–411 (2003)
5. Eastman, C.M., Jansen, B.J.: The appropriate (and inappropriate) use of query operators and their effect on web search results. Proceedings of the American Society for Information Science and Technology 41(1), 274–279 (2004)
6. Gospodnetić, O., Hatcher, E.: Lucene in Action. Manning Publications (2005)
7. Harman, D.K.: The TREC Test Collections. In: Voorhees and Harman [22], ch. 2, pp. 21–53
8. Hölscher, C., Strube, G.: Web search behavior of internet experts and newbies. Comput. Netw. 33, 337–346 (2000)
9. Jansen, B.J., Pooch, U.: A review of web searching studies and a framework for future research. J. Am. Soc. Inf. Sci. Technol. 52(3), 235–246 (2001)
10. Jansen, B.J., Spink, A., Saracevic, T.: Real life, real users, and real needs: a study and analysis of user queries on the web. Inf. Process. Manage. 36(2), 207–227 (2000)
11. Lucas, W., Topi, H.: Form and function: the impact of query term and operator usage on Web search results. J. Am. Soc. Inf. Sci. Technol. 53(2), 95–108 (2002)
12. Ogilvie, P., Callan, J.P.: Experiments Using the Lemur Toolkit. In: TREC 2001: Proceedings of the 9th Text REtrieval Conference. NIST, Gaithersburg (2001)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: OSIR 2006: Proceedings of ACM SIGIR 2006 Workshop on Open Source Information Retrieval (2006)
14. Palacio, D., Cabanac, G., Sallaberry, C., Hubert, G.: Measuring Effectiveness of Geographic IR Systems in Digital Libraries: Evaluation Framework and Case Study. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 340–351. Springer, Heidelberg (2010)
15. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library 27(6), 919–937 (2009)
16. Sanderson, M.: Test collection based evaluation of information retrieval systems. Found. Trends Inf. Retr. 4(4), 247–375 (2010)
17. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33(1), 6–12 (1999)
18. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the web: the public and their queries. J. Am. Soc. Inf. Sci. Technol. 52(3), 226–234 (2001)
19. Tukey, J.W.: Exploratory data analysis. Addison-Wesley, Reading (1977)
20. Voorhees, E.M., Harman, D.K.: Overview of the Seventh Text REtrieval Conference (TREC-7). In: TREC-7: Proceedings of the 7th Text REtrieval Conference, pp. 1–23 (1998)
21. Voorhees, E.M., Harman, D.K.: Overview of the Seventh Text REtrieval Conference (TREC-8). In: TREC-8: Proceedings of the 8th Text REtrieval Conference, pp. 1–23 (1999)
22. Voorhees, E.M., Harman, D.K.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge (2005)
23. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference, pp. 255–262. ACM, New York (2007)
24. Williamson, D.F., Parker, R.A., Kendrick, J.S.: The box plot: A simple visual method to interpret data. Ann. Intern. Med. 110(11), 916–921 (1989)

Evaluation Platform for Content-Based Image Retrieval Systems

Petra Budikova, Michal Batko, and Pavel Zezula

Masaryk University, Brno, Czech Republic

Abstract. In all subfields of information retrieval, test datasets and ground truth data are important tools for testing and comparison of new search methods. This is also reflected by the image retrieval community where several benchmarking activities have been created in past years. However, the number of available test collections is still rather small and the existing ones are often limited in size or accessible only to the participants of benchmarking competitions. In this work, we present a new freely-available large-scale dataset for evaluation of content-based image retrieval systems. The dataset consists of 20 million high-quality images with five visual descriptors and rich and systematic textual annotations, a set of 100 test query objects and a semi-automatically collected ground truth data verified by users. Furthermore, we provide services that enable exploitation and collaborative expansion of the ground truth.

Keywords: large-scale image dataset, visual and textual annotation, ground truth, collaboration service.

1 Introduction

Image search is a very attractive topic nowadays, as witnessed by the number of recent research papers and the rapid development of commercial image search systems. Still, a satisfactory solution seems to be yet a long way ahead due to a number of obstacles – the size of data, the semantic gap problem and a range of specific application issues. Many researchers are trying to overcome these problems using different ideas and approaches. To be able to pick out and develop the best ones, we need a common platform for testing and evaluation.

The lack of benchmarks for image search is a well-known problem. Although several freely-available test datasets exist, they do not cover all basic application areas. In particular, there is no large-scale dataset that could be used for evaluation of web image search, where scalability is one of the key issues. Therefore, we created a new extensive dataset and offer it hereby to other researchs.

Naturally, we need more than just the data to be able to compare the performance of various search methods. The other necessary part is the ground truth, i.e. the set of query objects and the respective sets of relevant result objects. Only then we can compare the precision, recall, and other metrics of the methods under examination. Unfortunately, creating the ground truth is a difficult

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 130–142, 2011.

task for any dataset, let alone the very large ones. As there is no objective automatic way of deciding the relevance of image with respect to a given query image, we need to ask people to do the tedious task. In case of large datasets, it is hardly possible to manually examine all the images in the dataset. Therefore, we adopt a different approach and describe a way of creating a partial ground truth by pre-selecting a candidate set of a reasonable size that is then examined and refined by humans.

In addition to the dataset itself and the ground truth for a hundred different query images we also offer two web-service tools. The first tool allows evaluating any search method against the ground truth. The second tool provides an interactive user interface that allows to collaboratively create a new ground truth for additional images.

2 Related Work

In the early days of image retrieval, the Corel dataset was the first collection to be used for evaluation. It provided over 68,000 images, organized into classes of about 100 images, each with roughly the same topic. However, this artificial and relatively small dataset is not satisfactory as a benchmark nowadays. We need to take into account different applications, the data they use (scope, size, metadata available, etc.) and the user information-retrieval requirements. Serious efforts for building a complex benchmarking platform appeared in [15]. The proposed methodology was to be realized by the Benchathlon[1] project, where research groups were meant to cooperate on creating the testing platform. Unfortunately, this project does not seem to be making any progress. Another analysis of image evaluation campaign can be found in [8]. It describes the background of ImageEVAL competition, which took place in 2006. However, the only repeated and successful benchmarking activity we know of is the ImageCLEF[2] competition which has been running since 2003. Each year, the organizers define various challenges, provide data and topics and evaluate the submitted results.

Nonetheless, even the ImageCLEF activities are limited by the availability of *benchmark inputs*, as defined in [12]: the data collection (documents), the queries (topics) and the ground truth (relevance judgements). We review these three issues in more detail in the following sections. We mainly focus on large, general-purpose datasets, leaving aside specialized collections such as medical images, arts collections, etc.

2.1 Image Databases

Gathering a large collection of image data is not a simple task due to the ownership and copyright issues. However, this can be overcome by using freely available web resources, such as the Flickr web gallery or Wikipedia. The following three datasets have been obtained this way. The first two have been composed to serve

[1] `http://www.benchathlon.net/`
[2] `http://www.imageclef.org/`

as the benchmarking sets and are used in the ImageCLEF competition. All of them provide both images and text metadata, but they differ in size, origins and scope of the metadata.

MIRFLICKR Collection. The MIRFLICKR collection[3] [9] consists of 1 million images (at the moment) downloaded from the social photography site Flickr. All images are available under a Creative Commons Attribution Licence. The images have been selected based on their high *interestingness* rating that is determined by factors such as where the click-throughs on the images are coming from, who comments on them, and whether they are marked as favorites. In addition, user-supplied Flickr tags, EXIF metadata and systematic image annotations are available. The visual descriptors provided are the MPEG-7 Edge Histogram and Homogeneous Texture descriptors, and the ISIS Group color descriptors.

Wikipedia Collection. The ImageCLEF 2010 Wikipedia collection[4] [18] extends the INEX MMWikipedia collection [19], which was created for the purpose of INEX evaluation campaign in 2007. Currently the collection consists of 237,434 Wikipedia images, their user-provided annotations, the Wikipedia articles that contain these images, and low-level visual features of these images. The collection was built to cover similar topics in English, German and French and it is based on the September 2009 Wikipedia dumps. Images are annotated in none, one or several languages. Image visual features include both local (bags of visual words) and global features (texture, color and edges). The collection is available for the participants of the ImageCLEF competition.

CoPhIR Image Set. The CoPhIR dataset[5] [1] with 106 million processed images is currently the largest collection available for scientific purposes. It consists of metadata extracted from the Flickr photo sharing system. For each image, the collection contains a thumbnail image, a link to a corresponding entry at the Flickr web site, user-specified metadata (title, GPS location, tags, comments, etc.) and five MPEG-7 visual descriptors (Scalable Color, Color Structure, Color Layout, Edge Histogram and Homogeneous Texture). Since the data are not supervised, some of them are of poor quality – blurred or too dark/light images, images with sparse and erroneous annotations, different languages used in annotations, etc. While this may cause worse performance of search methods, the collection provides a good model of a real-world data.

2.2 Topics

The common goal of all search systems is fulfilling user's information need. Therefore, the test search topics should simulate what a real user of the system would instantiate as usage scenario. Furthermore, the volume (number) and

[3] http://press.liacs.nl/mirflickr/
[4] http://www.imageclef.org/2010/wiki
[5] http://cophir.isti.cnr.it

diversity (variability) of the topics should cover the whole search domain and demonstrate statistical robustness of the results [12].

The usual ways of creating test search topics comprise a choice made by domain experts and an analysis of search system usage logs. In [18], the creation of topics for ImageCLEF 2010 Wikipedia Retrieval task is described in more detail. A candidate set of queries is derived from a search log file and topics from previous runs of the competition. From these, only such queries are accepted that have a sufficient number of relevant results in an organizers' search run.

Another issue is the query definition. Basically, there are three ways to go – query by example (image), query by text and query by both text and images. While query images are used in annotation tasks, they are not suitable for image retrieval since one image may often represent several concepts, while the imaginary user is only interested in one of them. Therefore, either complex text queries or images complemented by text are used in web-like image search tasks.

2.3 Ground Truth

The ground truth data is used to decide the relevance of a result provided by a search system. In an ideal case, the ground truth should contain an indicator of relevance for each object in the dataset and each search topic. The relevance can be either binary (relevant, irrelevant) or expressed as a level of relevance, e.g. as a percentage. The relevance is decided by human judges, preferably more than one for each object and topic to balance the subjectivity of opinions.

Clearly, creating such a ground truth is a laborious effort. When only a few people are involved, be them domain experts or lab members, providing exhaustive relevance judgements is only feasible for relatively small datasets. For the large ones, some approximations are usually employed. The one that is mostly used in the evaluation campaigns is called *pooling*: only those objects are judged that appear among the top n images of any of the results submitted by the competitors [18]. However, this results in a one-time ground truth that cannot be meaningfully reused for evaluation of different result sets.

Alternatively, expert annotations can be provided for each image, using a defined categorization. This is also a tedious work, but only needs to be done once for each object in the dataset. Afterwards, a ground truth for a given query can be obtained by judging only objects that have the relevant keywords in their annotation. This approach has been adapted by the supervisors of the MIRFLICKR dataset [9].

The only way to obtain an exhaustive ground truth for large datasets is by employing many people. However, it is not easy to find the necessary motivation. One possible approach is to invest a considerable amount of money and pay for each judgement. This approach was adapted to create the ImageNet database [6], where the Amazon Mechanical Turk platform was used to manually clean a large set of candidate images. Another method is shown in the TagCaptcha image annotation system [13], where the authors propose to obtain annotations via the widely used Captcha challenge-response tests.

Altogether, it is obvious that it is difficult to create the ground truth data. The evaluation campaigns such as ImageCLEF gather relevance judgements during the competitions but these are not public in order to prevent cheating in future competition runs. In consequence, there is a deficiency of ground truth data for testing outside the evaluation campaigns, which is definitely an obstacle in the development of new search methods.

3 Our Dataset

Our objective is to provide a dataset that will enable to test systems for large-scale searching in terms of results quality (precision), efficiency (search time) and scalability. The important aspects are therefore (1) the size of the dataset, (2) its scope, and (3) the type of data provided. As to the size, the datasets that are used for benchmarking nowadays range in volume from hundred thousand to millions of images. We believe that even larger datasets are necessary to test the efficiency of methods for web searching. Regarding the scope, we are interested in a real-world dataset since the performance of search mechanisms is influenced by the distribution of objects in the domain. Finally, recent research [5,10] indicates that the future of image searching seems to be in combining multiple modalities, typically visual features and text metadata describing the semantics. Therefore our dataset should contain at least these two modalities.

The Profimedia collection which we are offering to the research community satisfies all the discussed requirements. We obtained the image set from Profimedia[6], a web-site selling stock images produced by photographers from all over the world. The collection contains 20M high-quality images with rich and systematic annotations. For each image, we have extracted five MPEG7 [14] global visual descriptors recommended in [1]. Thus, each entry in the dataset consists of the following information:

- a thumbnail image;
- a link to the corresponding page on the Profimedia web-site;
- two types of image annotation: a title (typically 3 to 10 words) and keywords (about 20 keywords per image in average) mostly in English (about 95 %);
- five MPEG-7 visual descriptors extracted from the original image content: Scalable Color, Color Structure, Color Layout, Edge Histogram and Region Shape.

The dataset can be downloaded from **http://mufin.fi.muni.cz/profiset/** after registration and agreement to the usage terms. The data can be freely used for research purposes.

4 Query Topics

When selecting the topics, we had the following requirements in mind: the queries should reflect real users' needs, the topics should be diverse both in content and

[6] `http://www.profimedia.com/`

in complexity, and there should be enough relevant results for each test query in the dataset.

To achieve this we first created a set of candidate topics which comprised (1) popular queries from the search logs provided by Profimedia, and (2) several examples of queries that we know from experience to be either easy or difficult to process in content-based searching. Next, we run a top-30 query for each of the candidates using an aggregation of text and visual search (described in more detail in Section 5). Only the topics that had at least 10 relevant results were accepted into the final query set.

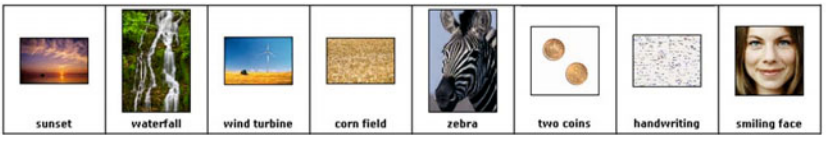


Fig. 1. Query object examples

The test set contains 100 topics, each of which is defined by a single query image and a few keywords (typically one or two). The following categories are represented by the topics: activity (5 queries), animal (8), art (6), body part (5), building (3), event (3), food (8), man-made objects (16), nature (16), people (12), place (9), plant (2), specific building (4), and vehicle (3). Several examples of the query objects are shown in Figure 1.

5 Partial Ground Truth

As stated earlier, a full ground truth should contain relevance evaluation for each topic-object pair but creating it for a large dataset is only feasible when a lot of people are employed. Unfortunately, we lack the resources required to do such a job, so we have used the pooling approach and our lab colleagues acted as the judges. We are aware of the fact that the pooling approach can miss relevant images, thus the provided data represent a *partial ground truth*. However, we tried to create it in such a way that it should cover the majority of relevant images and we also provide tools for expanding the ground truth when needed.

In the benchmarking competitions, the pool of candidate images for the evaluation is usually composed of the top n submitted results. We applied a similar technique but we have used a set of our own search methods implemented over the MESSIF framework [3] that provided the results. These methods were designed in such a way as to retrieve as many different relevant objects as possible. Our query topics consist of image and text, thus we can work with these two modalities and apply text-based retrieval, content-based retrieval or a combination of both. Furthermore, many preprocessing (query expansion) and postprocessing (ranking) methods have been proposed recently to improve the search

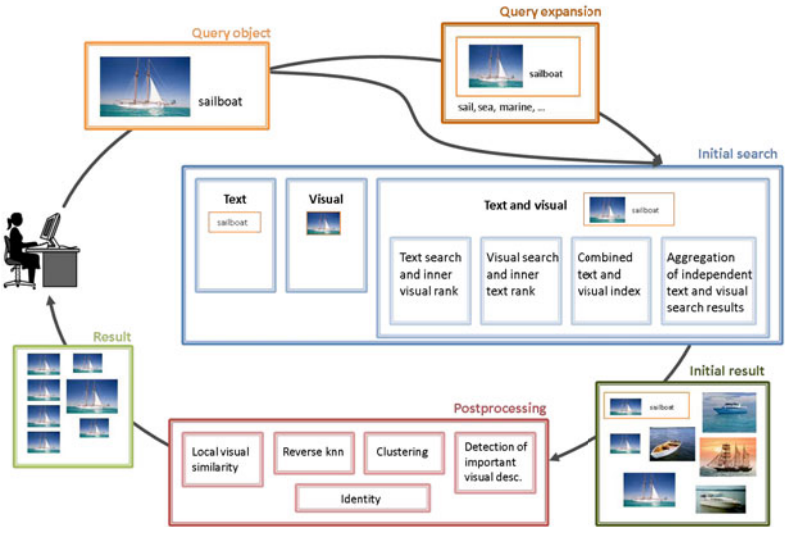


Fig. 2. The global search schema

efficiency. As illustrated in Figure 2, we treat the query evaluation as a three-phase process, where each of the phases can be realized in several ways. The search methods we used to create the pool of candidate images are then formed by various combinations of the individual techniques.

5.1 Query Expansion

The query expansion techniques [16] endeavour to automatically provide additional information to the query that will help to obtain better search results. Query expansion can be used to describe the user's information need more precisely (e.g. word sense disambiguation) or to overcome the gap between the query specification and the data available (e.g. automatic synonym expansion). For our test search methods, we chose the basic expansion technique that is often applied on short text queries. Using the WordNet [7] lexical database, we enriched the query with synonyms and hypernyms of the query keywords. This way, the relevant objects can be added into the candidate set even though their annotations are formulated differently.

5.2 Initial Search

In the initial search phase, the query (expanded or not) is submitted to a search method which processes it over the whole dataset and produces an initial result set. We adopt three types of searching: text-based retrieval, content-based (visual) retrieval and combined text-and-visual search. All the functions are implemented using the MESSIF framework [3] and the MUFIN [17] system.

Text-Based. The text search is executed as a classical *tf-idf* retrieval, only with different weights used for keywords from user, keywords in the query image title and keywords in the image annotation.

Content-Based. The content-based search is based on the five MPEG-7 descriptors available and the respective distance functions as defined in the MPEG-7 standard [14]. The individual distances are combined using a weighted sum.

Text-and-Visual. The combined search aggregates text and visual similarity. We use a weighted sum aggregation function with three different settings of the respective weights. The combined search can be implemented in several ways:

- Text search and inner visual rank: The text search is run on a full database but all objects relevant by the text criterion are ranked by combined text-and-visual similarity and only then the top ranking results are returned.
- Visual search and inner text rank: Same as previous, only vice-versa.
- Combined text and visual index: The search is evaluated over a metric index structure that combines the text similarity and content-based similarity.
- Independent text and visual search: The two search methods are run separately and the ranked lists are aggregated. In [2] we described how the aggregation can be done efficiently in a distributed environment.

5.3 Postprocessing

The philosophy of postprocessing is based on the fact that the search engine can provide a result set one or two orders of magnitude larger than required with nearly the same costs. Additional evaluations of similarity can be computed over this initial result that would be too expensive to process over the whole dataset [4]. In our experiments, we applied the following ranking functions:

- Identity: No ranking is applied, the top objects from the initial result are displayed to the user.
- Identification of important visual descriptors: The variance of visual descriptors is analyzed over the initial result set, the descriptors with low variance receive higher weights.
- Clustering: Objects that are more similar to the other objects in the initial result are ranked higher than the outliers.
- Reverse kNN: The rank of an object is given as the number of objects that are more similar to it than the query object.
- Local visual similarity: local similarity is evaluated using the SIFT features [11], the top ranking objects are shown to the user.

5.4 Relevance Judgement

Altogether with variable weights settings we created 140 search methods. For each query image, top-20 queries were evaluated by all methods and their merged

Fig. 3. The web interface for relevance evaluation

results were displayed in a web interface shown in Figure 3. The judges were asked to mark each object as *very good*, *acceptable*, or *irrelevant*, which we transformed into relevance levels of 100 %, 50 % and 0 %, respectively. Using the numerical values, we evaluated the final relevance as an average of collected judgements.

5.5 Statistical Evaluation

The ground truth data we obtained from our judges contain a considerable number of relevance evaluations which are a valuable resource for analysis of human perception of similarity. In this section, we present several observations concerning both the properties of our dataset and the human factor in the evaluation.

The evaluation was performed by 15 participants, most of them students, graduates, or researchers in IT. Out of the 100 queries, each got evaluated at least twice, the total number of evaluations being 222. With the average number of candidate objects per query topic being 578, we obtained a total of 128,141 evaluated topic-object-user triplets. The evaluation process took a month, the actual time invested in the judgements being about 100 hours.

As mentioned earlier, we compute the relevance of a result object as the average of all evaluations we have for it. We find it suitable to categorize objects into the following categories: *perfect* (average relevance 100 %), *good* (at least 50 %), *partially relevant* (more than 0 %), and *irrelevant*. For each query topic in our testbed, there were in average 105 perfect result objects, 223 good objects and 315 irrelevant ones. However, the number of objects in each category differed considerably between individual queries – the lowest number of perfect results was 5 and 11 objects had less than 20 perfect results. The lowest number of good results per query was 53. We can conclude that our set of topics is suitable for testing as there are enough relevant objects to be found and, at the same time, enough queries with various difficulty levels are present (difficulty being inversely proportional to the number of relevant objects contained in the dataset).

When evaluating the results, the judges were not given any instructions on what shall be considered relevant. Therefore, their classification of results reflects their individual understanding of similarity and their expectations of image search system performance. While this is known to be subjective and inconsistent

in different situations, all image retrieval systems are based on a tacit assumption that there exists some basic agreement in the individual opinions. Using our relevance evaluations, we can verify this assumption. The following table shows the percentage of identical evaluations, where all judges agreed on the (ir)relevance of a query object pair.

Number of evaluations	Identical evaluations	Unmatched (2 different)	Unmatched (3+ different)
2	80 %	20 %	–
3	70 %	27 %	3 %
4	73 %	21 %	6 %
5	65 %	20 %	15 %

For the sake of our ground truth it is also important to know whether two judgements (which we have for most queries) are sufficient to obtain a trustworthy relevance evaluation or whether more opinions are needed. Figure 4 shows how the percentage of objects with given relevance changes with the growing number of evaluations (we used the results with the most evaluations to obtain these graphs). We can observe that the results are quite stable, therefore the two judgements can be considered sufficient.

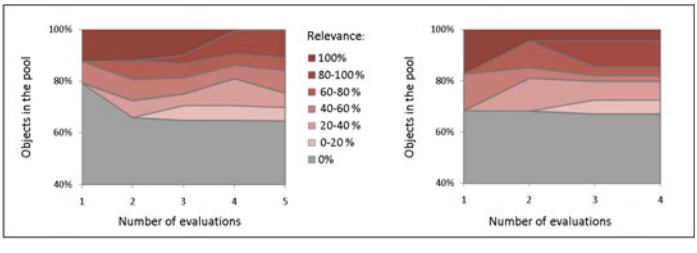


Fig. 4. The development of result evaluation

Finally, let us have a look on the methods we used to create the candidate pool. We employed a high number of combinations of search methods and post-processing techniques in order to discover as many relevant objects as possible. This approach has proved to be well suited as every combination did bring some relevant object to the results that was not found by any other method.

6 Provided Functionality and Extensibility

In order to offer the tools created during the preparation of the partial ground truth for the research community's benefit, we have designed two web-services. The first one simplifies the benchmarking of an external search method against the existing ground truth. The other one allows to add a new image to the query set and collaboratively evaluate its partial ground truth.

Due to space limitations we only explain the services in general here. More details, the specifications, and the access to the services can be found on the dataset page **http://mufin.fi.muni.cz/profiset/** in the Services section.

6.1 Evaluation of External Search Method

Researchers proposing new search methods for image retrieval systems are welcome to use the Profimedia dataset as a benchmark. By downloading the dataset, the query set, and the ground truth, they can compute their own statistics on the effectiveness of their method. However, since our ground truth is only partial – given that it was evaluated from a limited set of candidate objects (see Section 5) – the new proposed method can be penalized on images that are relevant to the query object but were not included in the candidate objects.

To overcome this problem, the results of the new method (just the identifiers of the images) can be uploaded to our service. The service then checks all the objects that were in the original *candidate* set from which the ground truth was computed and any new image is presented via the web-interface. The user is then able to judge whether each of the new objects is *very good*, *acceptable*, or *irrelevant* in the same way as when the previous partial ground truth was created. Afterwards, the statistics of the new method using the updated partial ground truth are displayed.

Any such addition to the existing partial ground truth is also stored in our database and immediately available for download. Thus, the partial ground truth is collaboratively extended whenever a new method is tested via our service.

6.2 Additional Query Images

Since our query set consists of a hundred images while the dataset contains 20 million images, we offer a service that allows to evaluate the partial ground truth for an additional query image. In order to do that, we need a candidate set of images and then user judgements of the relevance of the respective images (as explained in Section 5).

Our service thus allows to upload a new query image (or select an existing image from the Profimedia dataset using its identifier). Then one or more candidate sets can be uploaded, e.g. retrieved by some new search methods (as in the other service above). Finally, the system asks whether the candidate set should be expanded by our search methods. We provide options for selecting our text-based, content-based, or combined methods as explained in Section 5. Since the candidate set creation is a computationally intensive task, a job is scheduled in our university GRID[7]. It can take some time until the candidate set is ready for the user judgement, so the service notifies the user by email.

Then, the new query object is available for the user evaluation via a web interface as shown in Figure 3. When at least one evaluation is complete, the

[7] `http://www.metacentrum.cz/`

query is available in the query set with the new partial ground truth. The query is then also offered for additional evaluations to other users.

7 Conclusions

In this paper, we present a new freely-available large-scale dataset for evaluation of content-based image retrieval systems. The dataset consists of 20 million high-quality images with five visual descriptors and rich and systematic textual annotations. For this dataset, we have prepared a set of 100 test query images from various categories and collected a partial ground truth for each of them. The partial ground truth was human-judged from candidate sets generated by 140 search methods. To allow exploitation and collaborative expansion of the ground truth, we offer two public web-services. The data and services are accessible on the web page **http://mufin.fi.muni.cz/profiset/**.

Acknowledgments. This work has been partially supported by Brno PhD Talent Financial Aid and by the national research projects GAP 103/10/0886 and VF 20102014004. The hardware infrastructure was provided by the METACentrum under the programme LM 2010005.

References

1. Batko, M., Falchi, F., Lucchese, C., Novak, D., Perego, R., Rabitti, F., Sedmidubský, J., Zezula, P.: Building a web-scale image similarity search system. Multimedia Tools Appl. 47(3), 599–629 (2010)
2. Batko, M., Kohoutkova, P., Zezula, P.: Combining metric features in large collections. In: ICDE Workshops, pp. 370–377. IEEE Computer Society, Los Alamitos (2008)
3. Batko, M., Novak, D., Zezula, P.: MESSIF: Metric similarity search implementation framework. In: Thanos, C., Borri, F., Candela, L. (eds.) Digital Libraries: Research and Development. LNCS, vol. 4877, pp. 1–10. Springer, Heidelberg (2007)
4. Budikova, P., Batko, M., Zezula, P.: Similarity query postprocessing by ranking. In: 8th International Workshop on Adaptive Multimedia Retrieval (2010)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2) (2008)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR 2009 (2009)
7. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge (1998)
8. Fluhr, C., Moëllic, P.A., Hède, P.: Usage-oriented multimedia information retrieval technological evaluation. In: Multimedia Information Retrieval, pp. 301–306 (2006)
9. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proc. of the Multimedia Information Retrieval. ACM, New York (2008)
10. Jain, R., Sinha, P.: Content without context is meaningless. In: ACM Multimedia, pp. 1259–1268 (2010)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)

12. Marchand-Maillet, S., Worring, M.: Benchmarking image and video retrieval: an overview. In: Multimedia Information Retrieval, pp. 297–300 (2006)
13. Morrison, D., Marchand-Maillet, S., Bruno, E.: TagCaptcha: annotating images with CAPTCHAs. In: Proc. of the ACM Multimedia, pp. 1557–1558 (2010)
14. MPEG-7: Multimedia content description interfaces. Part 3: Visual. ISO/IEC 15938-3:2002 (2002)
15. Müller, H., Müller, W., Marchand-Maillet, S., Pun, T., Squire, D.M.: A framework for benchmarking in CBIR. Multimedia Tools Appl. 21(1), 55–73 (2003)
16. Natsev, A., Haubold, A., Tesic, J., Xie, L., Yan, R.: Semantic concept-based query expansion and re-ranking for multimedia retrieval. In: ACM Multimedia, pp. 991–1000 (2007)
17. Novak, D., Batko, M., Zezula, P.: Generic similarity search engine demonstrated by an image retrieval application. In: Proceedings of SIGIR 2009, p. 840 (2009)
18. Popescu, A., Tsikrika, T., Kludas, J.: Overview of the Wikipedia Retrieval Task at ImageCLEF 2010. In: CLEF (Notebook Papers/LABs/Workshops) (2010)
19. Westerveld, T., van Zwol, R.: The INEX 2006 multimedia track. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) INEX 2006. LNCS, vol. 4518, pp. 331–344. Springer, Heidelberg (2007)

Music Video Redundancy and Half-Life in YouTube

Matthias Prellwitz[1] and Michael L. Nelson[2]

[1] HTW Berlin, 10318 Berlin, Germany
[2] Old Dominion University, Norfolk VA 23508 USA

Abstract. YouTube is the largest, most popular video digital library in existence, and is quite possibly the most popular digital library regardless of format type. Furthermore, music videos are one of the primary applications of YouTube. Based on our experiences of linking to music videos in YouTube, we observed that while any single URI had a short half-life, music videos were always available at another URI. For this study we collected 1291 music videos and found that very few had zero or one copies in YouTube at any given time, and some had several thousand copies at any given time. Furthermore, individual URIs had a half-life of anywhere from 9 to 18 months, depending on the publication date and remaining commercial potential.

1 Introduction

YouTube is the leading video hosting service on the Web, with an *Alexa Traffic Rank* of three in November 2010 [1]. Similar services of have much lower ranks, e.g., dailymotion.com (Alexa Traffic Rank: 105), vimeo.com (166), myvideo.de (79). Due to its high popularity, YouTube is also a pioneer in struggling with copyright infringement as users upload (music) videos where they do not have the appropriate publishing rights, and copyright owners – mostly music record companies – identify these violations and petition YouTube to remove the offending videos. This ongoing publishing and removal of user-contributed music videos means that any particular music video likely has several functioning versions on YouTube at any given time, any specific URI is subject to be removed for a variety of reasons. For example, a copy of the music video "The Rolling Stones - Satisfaction" at `http://www.youtube.com/watch?v=214szPQBUYc` was removed on April 9, 2010 and the error message "This video is no longer available due to a copyright claim by ABKCO" is shown instead. Using YouTube's search functionality and querying the original video title in quotation marks returns 304 results (as of 12/10/2010) that could be understood as alternative available copies of the song.

We selected 1291 YouTube music videos from three different sources (U.S. Top 40, music blogs, and Rolling Stone's "500 Greatest Songs") and track the number of available copies of each music video as well as the half-life of individual URIs for a 10 week period. We quantify what we had observed anecdotally: individual copies of music videos are regularly uploaded and removed for a variety

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 143–150, 2011.

of reasons, but most music videos have multiple copies extant on YouTube at any given time.

Given its success, it is no surprise that YouTube has been studied and discussed innumerable times. Cheng et al. [4] characterizes the collection of videos at large, and also shows that music is the primary category (at 22.9%). Sharing on YouTube has been studied (e.g., [3]), as well as how people discover videos on YouTube [5]. YouTube's impact on popular culture and politics has been studied [8] [2], as well as ways for extracting the content and context for preservation purposes [9] [7]. But to the best of our knowledge, the half-life and redundancy of music videos in YouTube has not been studied.

2 YouTube HTTP Mechanics

A YouTube video URI is typically `http://www.youtube.com/watch?v=VIDEOID` where `VIDEOID` is a eleven character alpha-numeric identifier of the video. Dereferencing a YouTube URI returns a `200 OK` HTTP response code and the necessary HTML to play the video. Dereferencing a video's Atom feed at `http://gdata.youtube.com/feeds/api/videos/VIDEOID` returns a `200 OK` status code and the XML feed entry with metadata: video title, associated user, duration, tags, and optionally allowed or denied countries for seeing the video.

Once a video becomes unavailable, a `303 See Other` HTTP response code is returned[1] along with a `Location:` response header that gives a URI for an HTML page explaining why the video is unavailable: removed or blocked; classified as controversial; contains mature content; authentication required (i.e., private); captcha redirect[2]. We consider all but the last reason to indicate an unavailable video. Dereferencing these video's Atom feed URI results in an `403 Forbidden` HTTP response code and the metadata describing the video is now unavailable.

3 A Dataset of YouTube Music Videos

Since we were constructing a dataset of music videos that would be drawn from multiple sources (U.S. Top 40, music blogs, and Rolling Stone's "500 Greatest Songs"), we needed to be able to reliably gather metadata such as publication year and genre, for the music videos. For this purpose, we used Discogs[3], a free service providing detailed information about music releases with releases year, genre, track list, format, etc.

The first up to 20 search results identifying a musical "release" on Discogs were recorded and each HTML page was parsed out for publication year and genre. The release having the lowest publication date was chosen, and its publication year and genre(s) selected for the monitor set item.

[1] After this data reported in this experiment was taken, YouTube switched to conventional `404 Not Found` HTTP response codes for unavailable videos.

[2] See: `http://en.wikipedia.org/wiki/CAPTCHA`

[3] `http://www.discogs.com/`

The Top 40 songs of the US Singles Charts of September 25, 2010 [11] were chosen. Due to the relative homogeneity of the publication year and genre distribution of the 49 set items (there are more than 40 items to account for variations in artist and song metadata), the figure is omitted. Most of the songs were released in the current and previous years.

Three blogs from Blogger.com were selected that provide music reviews: `f-measure.blogspot.com` `youtube-music-videos.blogspot.com` `silaswillrock.blogspot.com`. The YouTube URIs were extracted from their Atom feeds, and the artist and title information was extracted from the YouTube HTML pages. For URIs not having this metadata, its video title was queried in quotation marks against Yahoo! search engine with the parameter "site:Last.fm", and the title and artist can be extracted from the structured URI returned by Last.fm In total 742 items were created. As expected, figure 1 shows a greater range of genres and publication years than the Top 40 dataset.

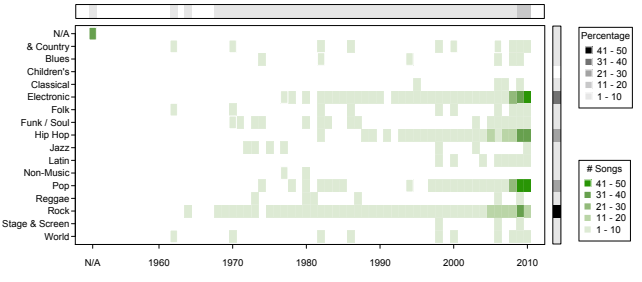


Fig. 1. Set distribution: Genre/Publication Year, Dataset: Music Blogs at blogspot.com, against Discogs. Number of items: 742.

Rolling Stone Magazine published a list of "The 500 Greatest Songs of All Time" [10], chosen by musicians, critics and industry figures. The song titles were mapped to YouTube URIs in the same manner as the Top 40 list. As expected from a list of this nature, Figure 2 shows a shift to older (i.e., "classic") songs as well as the genre roots of popular music (e.g., country, blues, funk/soul).

4 Experiment Methodology

We searched YouTube's GData API [6] with the query "ARTIST TITLE". The search is not a fielded search; it searches the entire page for the terms. The GData API allows retrieving of only the first 1,000 items, even if the total result size is larger (as indicated in a OpenSearch[4] element). We queried with the default chunk size of 25, up to the total result size or 1000 items.

After retrieving the whole list, the total result size value was stored according to the item with the current date. Afterwards, each feed item containing information about a video id was taken and processed: if a 'uri' element – identified

[4] `http://www.opensearch.org/`

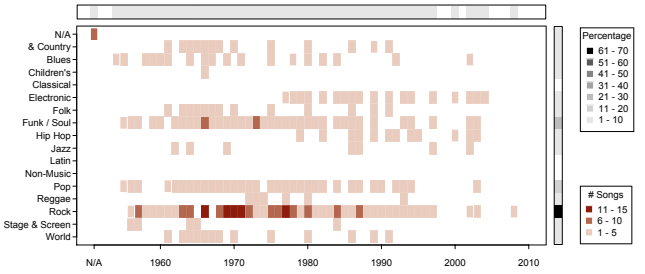


Fig. 2. Set distribution: Genre/Publication Year, Dataset: The 500 Greatest Songs of All Time by Rolling Stone, against Discogs. Number of items: 500.

by the video watch id – did not exist so far, a new record was inserted into the database with the static video properties: uploader (user), duration and publication date. If a 'uri' was not updated on the same day (as a 'uri' can be result of several monitor set items), its variable data – video title, rating, statistic, and comment counts, allowed/denied countries, and update date – were checked against a last dated database record in the corresponding table, if any, and a new tuple with the actual date was inserted according to the 'uri'. On change, the 'uri' is updated to the current date to prevent redundant processing if it appears in other search results.

For each feed item and transformed 'uri' object, its rank position is stored in relation to the search term, i.e. the 'item' object, with the current date. That allows keeping track of each URI over time with its (dis)appearance within the first up to 1,000 crawlable results, and its rank change in the feed.

A video URI can also fail to appear in any of the feed results of all items on a day. The reason might be that the rank of a video URI was lowered and its current position is greater than 1,000 that is not possible to observe due to the GData limitation. Another reason can be that a video is no longer available. To check the state of these videos, a cron job, starting after the previous one terminated, takes all 'uri' objects not updated on the current date and dereferences each URI. Receiving a `200 OK` HTTP status, the 'uri' record is just updated with the current date in its 'last crawled' property. With a `303 See Other` redirect, the 'uri' object is set to inactive with the current date, and the reason (parsed from the HTML page) is stored.

5 Results

The Top 40 dataset was initiated on October 1, 2010, Top 500 dataset on November 7, 2010, and the music blogs dataset on November 13, 2010. For the total of 1,291 monitor set items, 902,869 YouTube video URIs were discovered by December 12, 2010.

Figure 3 shows the Top 40 dataset with its items in descending median result size. The lines to each item indicate the item's maximum and minimum total

result size returned over the observation period. For example, the first item with the highest median of results (83,298) has a variation over time from a minimum of 43,945 to a maximum of 123,239 total results given. The last item with the lowest median of 35 results developed a minimum of 26 and a maximum of 66 over time. There was never a time where no videos were available for any song in this collection.

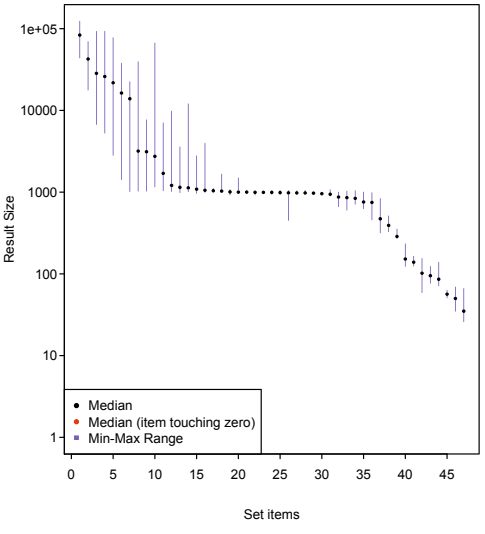


Fig. 3. Total Result Size Dataset: Top 40 US Single Charts of 9/25/10

A similar high variation of total result sizes of 1,000 or more also applies to the music blog dataset (Figure 4). For example, the first item with a median size of 255,581 varies over the given time up to 31,817 between maximum and minimum total result size. The range of items touching zero with their median or minimum result size are shown in detail in the sub plot: from 50 items that retrieved at least once zero results, 44 items never had a result returned. This could be because the video title does not accurately describe the song.

Finally, the Rolling Stone dataset also starts with a high total result size of median 145,076 for the first item and nearly 80% having a median of at least 100 copies. (Figure 5). Only 1%, five items retrieved zero results at least once, and four never received a result, which might be due to the accurate complete artist and song title information, as well as the popularity of songs on the list.

We kept a list of all unique URIs discovered from the daily result sets described above so we could trace how long individual URIs persisted. Rooted from that set, the daily availability of those URIs was measured over time. Normalized and aggregated monitor set-wise, figure 6 shows the removal rate over the observation period. Due to the different start dates of the three monitor sets, different durations are present. A continuous unavailability rate can be concluded for each set, visualized by its median values. Furthermore, the plot shows a higher rate

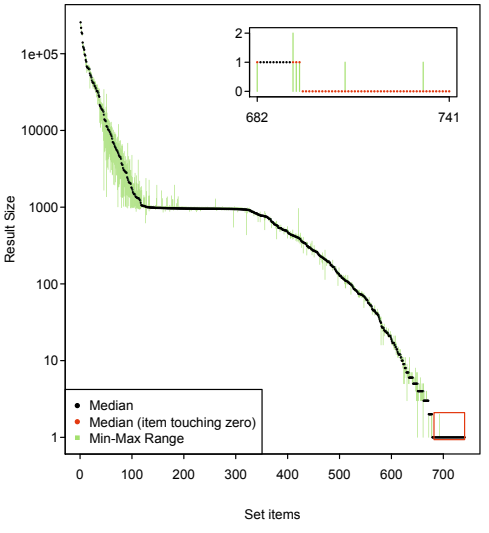


Fig. 4. Total Result Size Dataset: Music Blogs at blogspot.com

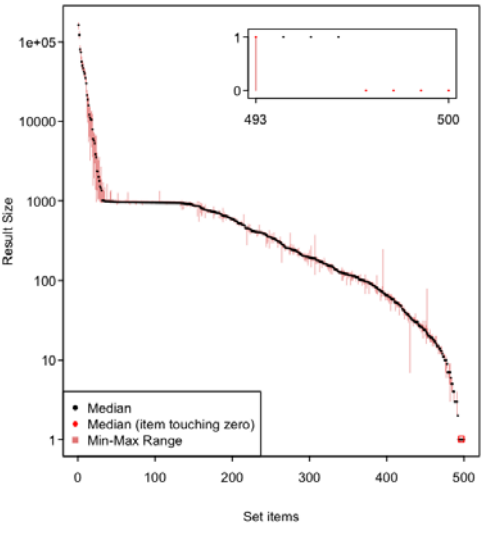


Fig. 5. Total Result Size Dataset: The Top 500 Songs of All Time

of removal for the Top 40 set, presumably because of their current economic potential they are more actively policied by their copyright holders.

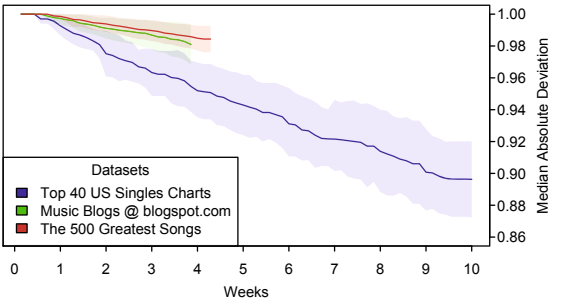


Fig. 6. Unavailable URIs

Applying linear regression to each monitor set and predicting its progress is an interesting aspect of the half-life of each collection. Figure 7 shows the monitor set-wise aggregated regression.

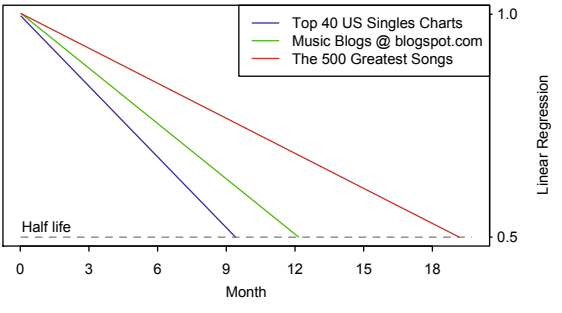


Fig. 7. Predicted Half-Life of Collection

Nearly half of the video removals (48.8%) were a result of third-party claims by copy right holders. For 23.3% of the removed videos, YouTube removed a video or discontinued the user account, e.g. due to violations against one of its policies or its terms of service. For only 13.2% did users voluntarily remove the video or close their account. The remaining group summarizes observed crawling errors or status changes of video, e.g. the user set a video to private.

6 Summary

We collected 1291 music videos from three collections: a Top 40 U.S. Singles chart (49 videos), a series of blogs that link to music videos (742 videos), and the Rolling Stone list of the "Top 500 Greatest Songs of All Time" (500 videos). We have shown that for music videos in YouTube, one can expect the URI for any given video to be short lived, with half-lives of 9 (Top 40) to 18 months

(Greatest 500) calculated from our datasets. This suggests that the more recent and popular the song (and thus, the more economic potential it represents), the more likely there will be thousands of copies at any given time, as well as copyright holders aggressively requesting their takedown. YouTube provides their Content ID[5] software suite as a method to help copyright owners to identify when their intellectual property is being used (and optionally, to monetize its use). Despite its use, although individual URIs come and go quite frequently, the music video persists, in aggregate, quite well in YouTube.

References

1. Alexa: Youtube.com Site Info (2010), http://www.alexa.com/siteinfo/youtube.com (last checked: December 11, 2010)
2. Capra, R.G., Lee, C.A., Marchionini, G., Russell, T., Shah, C., Stutzman, F.: Selection and context scoping for digital video collections: an investigation of youtube and blogs. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 211–220 (2008)
3. Cheng, X., Dale, C., Liu, J.: Understanding the characteristics of internet short video sharing: YouTube as a case study. Technical Report Arxiv preprint arXiv:0707.3670 (2007)
4. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: 16th International Workshop on Quality of Service, IWQoS 2008, pp. 229–238 (2008)
5. Cunningham, S.J., Nichols, D.M.: How people find videos. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 201–210 (2008)
6. Google: Developer's Guide: Data API Protocol - API Query Parameters - YouTube APIs and Tools - Google Code (2010), https://code.google.com/apis/youtube/2.0/developers_guide_protocol_api_query_parameters.html#qsp (last checked: December 12, 2010)
7. Marchionini, G., Shah, C., Lee, C.A., Capra, R.: Query parameters for harvesting digital video and associated contextual information. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 77–86. ACM, New York (2009)
8. Sayre, B., Bode, L., Shah, D., Wilcox, D., Shah, C.: Agenda Setting in a Digital Age: Tracking Attention to California Proposition 8 in Social Media, Online News and Conventional News. Policy & Internet 2(2), 2 (2010)
9. Shah, C.: Tubekit: a query-based youtube crawling toolkit. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 433–433 (2008)
10. The Rolling Stones Magazine. The RS 500 Greatest Songs of All Time: Rolling Stone (2004), http://web.archive.org/web/20080622145429/www.rollingstone.com/news/coverstory/500songs (last checked: November 13, 2010)
11. top40-charts.com. USA Singles Top 40 @ Top40-Charts.com (2010), http://top40-charts.com/chart.php?cid=27&date=2010-09-25 (last checked: December 13, 2010)

[5] http://www.youtube.com/t/contentid

Linguistic and Semantic Representation of the Thompson's Motif-Index of Folk-Literature

Thierry Declerck[1,2] and Piroska Lendvai[3]

[1] Austrian Academy of Sciences, ICLTT,
Sonnenfelsgasse 19/8, 1010 Wien, Austria
[2] DFKI GmbH, Language Technology Lab,
Stuhlsatzenhausweg, 3, 66123 Saarbrücken, Germany
[3] Hungarian Academy of Sciences, Research Institute for Linguistics,
Benczúr u. 33., 1068 Budapest, Hungary

Abstract. We present on-going work on the linguistic and semantic processing of the labels of the Thompson's Motif-Index of Folk-Literature, which has been proposed by Stith Thompson for the classification of narrative elements in folk-literature. We automatically extracted the labels of an on-line version of the Index, and wrote specialised grammars for providing for a multi-layer linguistic annotation of them. We are currently working on enriching the linguistically annotated labels with semantic classes and relations, allowing for a better access to the content of the Index. With this resource, we expect to be able to semi-automatically annotate digitised literary works at the sub-document level by means of automatically comparing the annotated Index with the results of text processing tools applied to those works, and so contribute to a better inter-textual interlinking and understanding of related works in the folk-literature, offering a new way of semantically accessing digital libraries.

1 Introduction

The modelling and processing of motifs in text poses significant challenges for research in several disciplines. A motif is a cultural element that keeps recurring within or across artefacts - e.g. in film, music, but also in folk-literature -, by means of which an idea or theme is conveyed.[1] Motifs can be seen as semantically compact structures, representing cognitively complex notions, while they are mostly expressed in a wide and loose variety of lexical and syntactic realisations in folk-literature.

Linguistic features of motifs have so far not been systematically investigated, their computational modelling and classification is thus unresolved. Due to this technological gap, a large amount of cultural heritage collections of which motifs are typical constructive units still can only be manually indexed, mostly only at document level, which significantly limits access to these resources.

[1] Some random examples for motifs in folktales are e.g. the cruel stepmother, the poor girl who was chosen as wife in preference to a rich one, or a supernatural who substitutes the hero in a tournament.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 151–158, 2011.

In this paper we describe an approach for an automated linguistic and semantic analysis of the textual labels of a large resource, which is used by many researchers in the broad folktale domain for the classification of the literary works in terms of motifs: the Thompson's Motif Index of Folk-Literature ([1]). Linguistic and semantic normalisation of the textual expressions of motifs in those labels is accomplished. Those normalised labels can subsequently be used for detecting and indexing lexical and syntactic variants of motifs in linguistically and semantically processed folktale texts, also at sentential or even clausal levels. Assuming an acceptable quality of this indexing procedure, which can also be ensured by manual post-processing of the folktale texts, we expect this approach to open new ways for accessing digitized folktales and for establishing inter-textual relationships.

In this paper we first introduce the relevant aspects of the Motif Index of Folk-Literature (TMI) for our work. We then describe our methodology for the linguistic and semantic annotation of the labels of TMI. In a next section we show how this enriched classification resources can be used for the indexation (or annotation) of linguistically and semantically processed folktales. Finally, we draw some conclusions and point to future work.

2 Thompson's Motif Index of Folk-Literature

The Thompson's Motif Index of Folk-Literature (TMI, see [1]) is a hierarchically structured catalogue, grouping and identifying many thousands of motifs from folktales, myths, and other narrative genres from around the world. Contrary to type indexes that cluster tales based on their common plot elements[2], the TMI catalogue focuses on motifs that emphasise ideas or themes, and as such can be regarded to promote micro-level conceptual analysis in folk narratives.

Each motif-entry in the catalogue consists of a combination of an uppercase letter and digits, followed by a textual label. Uppercase letters not followed by a digit indicate a main category, covering a series of related and hierarchically ordered motifs. For example, "K. *Deceptions*" is a category, which spans among others over the particular motifs "K0-K99. *Contests won by deception*", one of which being "K3. *Substitute in contest*". Dozens of subtypes are assigned to this single motif; these catalogue descriptions, or labels, are short phrases such as"*Supernatural substitute in tournament for pious warrior*", "*Wise man disguised as monk beats learned heretic in debate*", and so on.

Such text segments play several roles according to various aspects that operate complex cognitive processes underlying narration; identifying and cross-linking these is a technologically non-trivial task. Regarding the TMI labels as the object of linguistic processing and annotation can benefit the disciplines of Linguistics, Folklore, Language Technology, and Digital Libraries, based on which we (a) propose a method for offline indexing of narrative resources, and (b) illustrate how it enables query expansion on cultural heritage texts.

[2] E.g. the Aarne-Thompson-Uther (ATU) classification system, cf. for example `http://oaks.nvg.org/folktale-types.html`

The TMI lists 23 main categories[3] and provides a hierarchical structure of motifs. Our objective is pinpoint linguistic and perhaps cognitive properties and elements shared between sets of motifs which enables them to connect to one another. Segmentation, i.e. addressing granularity of meaningful units in several interdependent linguistic levels is thus a core phenomenon (cf. [2]). Importantly, by utilising the TMI, the problematic issue of identifying motif boundaries can to a large extent be circumvented, since the TMI consist of pre-segmented, semantically aligned phrases. Nonetheless, semantically related motifs sometimes cannot be found in the TMI, since they are cognitively or typologically grouped under unlinked motifs: e.g. there are several further occurrences of terms that include the component 'substitute' in the TMI but are located elsewhere than under node K, such as 'Substituted eyes' (under E780.1. Vital body: kills attacking enemies), 'Test of friendship: substitute as murderer' (under H1558. Tests of friendship).

Our approach will resolve such suboptimal, distributed representation of components of lexically/semantically related motifs. It will be possible for users of the TMI, as well as for linguists, to query lexical realisations of concepts occurring in motifs. For example, that 'contest' can be a 'tournament' or a 'debate', animate participants of the substitution can be e.g. wise man' (as substituted person, at the same time grammatical Agent, as well as the Hero of the tale) and 'learned heretic' (as opponent in the contest, as well as grammatical Patient, at the same time possibly the Villain or another prototypical character), 'warrior' and 'monk' are forms of disguise in the act of deception, 'disguise' and 'substitute' are verbs conveying the act of deception, etc.

3 Processing and Linguistic Annotation of TMI

We first pre-processed an on-line version of TMI[4], extracting only the codes of the index and the labels. We obtained a flat list of nearly 36,000 phrases of the following form:

```
0 :: Creator
2 :: Sun-god as creator A1.1.
3 :: Grandfather as creator A1.2.
8 :: First human pair as creators A2.2.
29 :: Human creator A15.
30 :: Female creator A15.1.
34 :: Old man with staff as creator A15.3.1.
35 :: Artisan as creator A15.4.
36 :: Potter as creator A15.4.1.
46 :: Sun and moon (man and wife) as creators A19.1.
49 :: Woman who fell from the sky A21.1.
```

[3] e.g. Animal Motifs, Magic, the Dead, Marvels, Tests, the Wise and the Foolish, Deceptions, Reversals of Fortune.

[4] See `http://www.ruthenia.ru/folklore/thompson/index.htm`

```
50 :: Old man from sky as creator A21.2.
53 :: Creator from below A25.
54 :: Creator emerges from lake A25.1.
69 :: Bee as God's spy A33.3.1.
70 :: Other animal companions of creator A33.4.
...
```

Many labels consist of a simple noun phrase ("Human creator"), but some incorporate complements, such as prepositional phrases ("Old man with staff as creator") and even verbs that produce sentences ("Creator emerges from lake").

We submitted the textual labels to a range of natural language processing tasks, including lemmatisation, part-of-speech and morphological tagging, detection of constituency structures and dependency relations, to be further enriched with argument structures, lexical-syntactical correspondences, and information about lexical relations from WordNet[5].

We opted for using the NooJ Linguistic Development Environment[6] for the (first steps) of implementation of the automated linguistic and semantic annotation of the labels of TMI. NooJ is supporting an easily configurable pipeline of linguistic processing tasks, which we describe in the following subsections.

3.1 Lexical Analysis

In a first step, we provide for a lexical analysis of the words included in the labels.

```
woman,N+Nb=s+Distribution=Hum
who,PRO+Distribution=RelQ
fell,fall,V+Tense=PT+Pers=3+Nb=s
from,PREP
the,DET
sky,N+Nb=s
```

In this analysis of the short sentence "Woman who fell from the sky" (TMI number: A.21.1), we can see the results of the process of lemmatisation (the verb form *fell* is reduced to its lemma *fall*. Part-of-Speech information has been provided (*woman* being a N(oun), *fell* being a V(erb), *the* being an Det(erminer), *who* being a Pro(noun) and *from* being a Pre(position). Morphological analysis also took place: *woman* is marked as singular (Nb=s), the verb *fell* is marked as being in the past participle tense (Tense=PT) and third person singular (Pers=3+Nb=s)[7]. We also note that some semantic information is available at the lexical level: woman is classified as being Hum(an).

[5] `http://wordnet.princeton.edu/`

[6] See `http://www.nooj4nlp.net/pages/nooj.html` or [3]

[7] A simplified representation his provided here. The verb form *fell* is morphologically highly ambiguous, and it could be marked with other Person or Number properties. Disambiguation takes place at a later processing level in NooJ.

3.2 Syntactic Analysis

We wrote specialised rules for performing the syntactic analysis of the textual content of the labels of TMI. This tailored solution is motivated by the relatively low number of linguistic patterns used in this resource, so that implementation of hand-crafted syntactic rules is not time consuming. An alternative would have been to manually annotate a relevant number of labels and train a parser for the automatic syntactic annotation of all the labels. We show below the resulting syntactic annotation of one motif:

```
<NP>
  <NP>
    <HEAD><REFOF XREF="396.2">Woman</HEAD>
  </NP>
  <SENT><RELCLAUSE>
    <SUBJ><XREF>who</XREF></SUBJ>
    <PRED>fell</PRED>
    <PP><PPOBJ>from
      <NP>
        <SPEC>the</SPEC>
        <HEAD>sky</HEAD>
      </NP>
    </PPOBJ></PP>
  </RELCLAUSE></SENT>
</NP>
```

We provide syntactic information on both the constituency (phrasal grouping of word forms) and the dependency relations between (groups of) word forms. For example, "woman" and "the sky" are marked as a constituent of type NP. At the dependency level, *woman* is marked as the Subj(ect) of the Pred(icate) *fell*. The detection of this Subj-Pred relation is possible only on the basis of an earlier round of computing the co-reference relation between *woman* and *who*, which we marked with the XML element "XREF". We also mark the dependency relation between the word forms within a phrasal constituent.[8]

3.3 Semantic Analysis

As mentioned above in section 3.1, NooJ offers the possibility to mark up lexicon entries with semantic information, like Hum(an). One can use for this any kind of semantic resource. Our semantic classification of lexical entries remains at a

[8] E.g. *sky* is the HEAD of the NP, whereas *the* is the corresponding Spec(ifier). In case we would have an adjective within an NP, like "the divine sky", the adjective would be marked as a MOD(ifier).

generic level: *Hum* (people and human-like creatures); *Inst* (groups and institutions); *Abstr* (abstracta); *Concr* (concreta); *Geogr* (geographical names); *Tier* (animals); *Anim* (life forms, non-human and non-animal)[9].

For the domain or application specific semantic annotation we opted for writing specialised semantic grammars, which can state fro example that a "Creator" can be either a human or an animal, and so on. We can also further restrict the potential animals to be just those named in the chapter "A" of TMI (the chapter about creation). The (simplified) grammar rule looks like this:

```
Main = <E>/<CREATOR :Creator <E>/> ;

Creator = ( :Animal | :Human ) ;
Human = <N+Hum> ;
Animal = <N+Tier> ;
# or alternative :
# Animal = (beast|insect|eagle|...) ;
```

This rule stipulates that a creator can be either a human or an animal. "Human" is defined on the base of the lexicon encoding, and thus all the labels containing a lemma of word forms having this semantic attribute will be displayed as a result of a query for "Creator". As an alternative, one can restrict the annotation to more specific entities, and list for example the animals that are indeed named in the TMI as potential creator.

4 Access to Properties of Motifs

All annotations provided so far can be queried by the NooJ system, enabling to search for e.g. humans in the subject position of a sentence. This allows to query for items playing different roles in a motif: e.g. a `woman` falling from the sky (as a creator), or as the mother of a hero. As a first step for judging the benefits of our approach, we compared our work with an on-line querying facility that is accessing the same data[10], an effort using only string matching between the query of the user and the the labels of the TMI motifs.

4.1 Querying TMI with a String-Based Matching Search Engine

We tested the search engine by querying the words "creator / creators" and "woman / women". The query "creator" returns all labels containing the words "creator" and "creators". Querying the word "creators" returns all labels containing "creators". More restricted is the search with the word pair "woman / women". Querying TMI with "woman" does not result in labels containing the plural form , due to the irregular plural form. The string-based matching algorithm obtains low recall of naive user's queries of entities.

[9] This is the semantic classification proposed by Ralph Mueller for his German NooJ module, see `http://www.nooj4nlp.net/pages/german.html`

[10] The URL of this service is `http://storysearch.symbolicstudies.org`

4.2 Querying the Linguistically and Semantically Annotated TMI with the NooJ Tools

Compared to the search interface described above, the NooJ querying facilities do not deal only with the strings of the word forms, but with all types of provided annotations, being at the string, lexical, syntactic or semantic level, as well as their combination Some examples are given below:

- Query on ' '<women>" (the angle brackets mean searching for the lemma) returns all the occurrences of labels containing both "woman" and "women"
- Search for all nouns in plural form (with the query: <N+Nb=p>.) Due to in-built lexicons in NooJ not only all labels with a noun bearing a plural marker are returned, but also labels containing e.g. "pair", a sematically plural word, as in the motif "First human pair as creators (A2.2)"
- Search for NPs, denoting a human, in subject position retrieves e.g. woman, man, sons, old man, etc. when in a syntactic subject position
- Search for semantic categories: <CREATOR> is now assigned to labels containing words such as 'god', 'dragon', 'eagle', and so on, based on the taxonomic structure of the TMI.

5 Linguistic and Semantic Analysis of Folktale Texts

As a first test, we submitted a German version of the Russian folktale *The Story of King Frost*[11] to our NooJ analysis module. Additionally to the TMI annotated labels, we use an in-house developed family ontology[12], which allows us to detect semantic relations, like for example that the two girls in the story are also stepdaughters, etc. We can in this case automatically detect two TMI motifs: "Tasks assigned by stepmother" (H934.3) and "Evil stepmother orders stepdaughter to be killed" (S322.4.2), with the additional information of who is receiving the order (in this case, the father of the good girl). But we also discovered that we could automatically detect Aarne-Thompson types[13], like type 480 "the kind and the unkind girls", type 312A "the saved girl". This finding is guiding our future work, which will consist in also providing for linguistic and semantic annotation of the Aarne-Thompson-Uther (ATU) classification system, and to combine both approaches into one semantic annotation frame for folktales.

6 Conclusion and Future Work

We have described an approach for the automated linguistic and semantic annotation of an important terminological and taxonomic resource in the broad field of folklore research: the labels of the Thompson's Motif-Index of Folk-Literature (TMI). Due to the limited number of linguistic patterns used in this resource,

[11] See `http://www.mythfolklore.net/andrewlang/017.htm`

[12] We would like to thank Nikolina Koleva for her work on this resource.

[13] `http://oaks.nvg.org/folktale-types.html`

we could rapidly develop specialised grammars that cover a large spectrum of TMI. Our main expectation is that this rich annotation of the TMI resource, offering linguistic and semantic frames, will guide the semi-automatic processing and domain specific indexing of folk-literature. The potential of automatically recognising and annotating motifs in text brings benefits to research programs in several disciplines addressing narrative discourse elements, and add a new dimension of creating advanced finding aids in Digital Libraries.

In ongoing work we are investigating the utility of the taxonomic structure of the TMI for incorporating cognitive knowledge organisation in queries, and its representation as subject-predicate-object expressions (RDF form), to not only automatise finding information, but its understanding as well. We are looking for extending the TMI to other languages, drawing on work done within the European project Monnet[14], which is dedicated to offer translations of knowledge systems. Monnet is for antoher reason central for us, since the project also delivers a representation schema for natural language expressions contained in taxonomies and ontologies. With this schema we hope to ba able to ensure interoperability of labels analysis of TMI and, for example, ATU. Results of our work will be distributed via the web page of the AMICUS project[15].

References

1. Thompson, S.: Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Revised and enlarged edition. Indiana University Press, Bloomington (1955-1958)
2. Lendvai, P.: Granularity Perspectives on Modeling Humanities Concepts. In: Darányi, S., Lendvai, P. (eds.) First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, Vienna, Austria. University of Szeged, Hungary (2010)
3. Silberztein, M.: Complex Annotations with NooJ. In: Blanco, X., Silberztein, M. (eds.) Proceedings of the 2007 International NooJ Conference, pp. 214–227 (2008)
4. Declerck, T., Scheidel, A.: An Information Extraction Approach to the Semantic Annotation of Folktales. In: Darányi, S., Lendvai, P. (eds.) First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, Vienna, Austria. University of Szeged, Hungary (2010)
5. Declerck, T., Scheidel, A., Lendvai, P.: Proppian Content Descriptors in an Augmented Annotation Schema for Fairy Tales. In: Sporleder, C., Zervanou, K. (eds.) Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Lisbon, Portugal. IOS Press, Amsterdam (2010); European Coordinating Committee for Artificial Intelligence – ECCAI 2010

[14] See `http://www.monnet-project.eu`

[15] AMICUS (Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts) is a network funded by NWO, see `http://amicus.uvt.nl`

WPv4: A Re-imagined Walden's Paths to Support Diverse User Communities

Paul Logasa Bogen II, Daniel Pogue, Faryaneh Poursardar, Yuangling Li, Richard Furuta, and Frank Shipman

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

Abstract. The Walden's Paths Project, as part of our philosophy of continual evaluation, actively seeks out user communities who may find our system to be of interest. In the past few years we noticed a recurring trend of user issues, needs, and sought-after features. In order to better support our users, we initiated a redesign of Walden's Paths that not only solves these problems, but enables us to perform more rapid prototyping and experimentation of new features and interfaces. In order to accomplish these goals, we have created a web service that handles the storage, modification, and representation of our path data structures. This service is completely isolated from user interface layers, allowing many different interface designs to be implemented on top of the basic Walden's Paths data structures. We also present several prototype interfaces – Marginalia, CoWPaths, Walden's Drupal, PathCompiler v2, mWalden – that represent new areas in which we believe our ideas can be applied such as collaborative work, location-aware services, large educational databases, offline presentation, and mobile computing.

Keywords: Walden's Paths, Collaborative Authoring, Web Services, Computer-Aided Education, Narrative Structures.

1 Introduction

Walden's Paths is a tool for building linear narrative structures from found resources on the web. A user, such as a teacher or a peer group of students, identifies web resources that they wish to use to illustrate a concept, tell a story, or provide as a reference collection. These pages are then ordered in a linear path. Walden's Paths can be viewed as a digital library of annotated indices of the web. Users then can provide annotations to help contextualize the pages, draw attention to particular passages, summarize the content, or inform readers of questions that they should be able to answer after reading the resource. We draw upon Bush's concept of associative trails that sit above the text building a new hypertextual lattice on top. When Walden's Paths was first conceived of in the mid-1990s, the Internet was a far different place. Pages were primarily static affairs and the concept of a web

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 159–168, 2011.

application was not yet fully realized. The original Walden's Paths system was built with this world in mind. While new ideas and features had entered the project both in the late-1990s with support for online-path creation in v2 [11], and in the early 2000s with more experimental features such as dynamic paths [3] and branching paths in v3 [4], the structure of the system had remained unchanged. During the era of Walden's Paths v3, experimentation was done primarily through branching of the code base. This led to pollution of the trunk codebase with experimental code and the loss of features, as branches were never re-incorporated into the base. Lastly, as the v3 codebase approached its tenth year, we began to see browser security features and the ubiquity of the interactive web wreak havoc on our system. Therefore, we decided to create WPv4, a re-imagined Walden's Paths for the modern web.

Since our last major, other projects have begun exploring paths. One project, HATS, produced a browser plug-in that allowed a flat-file representation of a pre-authored paths to be downloaded to the user and then navigated [7]. Others like, WebPath, appear as a faithful reimplementation of Walden's Paths for specific purposes, in this case building paths from resources inside LotusNotes [9]. Finally mSpace, abandons the web as it is at the surface and instead uses the deeper semantic web to build path-like structures without the *in situ* focus of Walden's Paths [10].

The remainder of this paper begins with discussion of a series of informal evaluations where we gathered feedback and observed the behaviors of Walden's Paths users; followed with a review of our past metaphors and a presentation of a new metaphor that describes WPv4. After this we introduce the WPv4 and discuss the current interface prototypes. Finally we present our conclusions and future plans.

2 Observations and Feedback

The Walden's Path Project regularly seeks out new users. When a new user group is identified, we provide them a quick tutorial and then observe their first experiences creating a path, provide help as requested, and solicit feedback on their experiences. These informal evaluations have revealed a number of common user issues, needs, and desired features. This section discusses the observations and feedback we have obtained from three groups we have worked with.

2.1 Atmospheric Sciences

The Department of Atmospheric Sciences at Texas A&M University teaches a basic course on Atmospheric Science both as a core science service course and as an introduction to new majors in Atmospheric Science. The course is offered in three different versions, one without a lab, one with a lab for non-majors, and one with a lab for majors. Every semester multiple teaching assistants for each of the three versions collaboratively work on the materials for the semester both for in-class usage and supplemental self-guided instruction.

The Walden's Paths project became involved with the preparation of self-guided tutorials on atmospheric science topics such as cyclone formation and weather radar. Creation of these tutorials illustrated previously unseen work practices with paths. Therefore, we identified the practices as areas where we could improve the user's workflow with enhancements to Walden's Path.

Due to the collaborative nature of material creation, users wished to share paths amongst each other. Instead of using the features provided, that allowed users to save and then load into other users' account, they instead decided to create a shared account that was used as a group space. In order to deal with differences between the three versions of the class, paths were duplicated in the group account and then modified for the alternate versions of the class.

When a new semester arrived, the paths were revisited by the new set of teaching assistants and revised to reflect changes on the web and experiences from past semesters. These scenarios established a need for Walden's Paths to support group-based collaboration, path extension, and path branching.

2.2 Pre-Service Teachers

The Ensemble Project is a NSF/NSDL-funded multi-university effort to provide access to collections, support communities, and develop tools to support computing education through an online portal. Through the Ensemble Project, the Walden's Paths project became involved with a faculty member in the Department of Teaching, Learning, and Culture at Texas A&M. The faculty member became interested in presenting Walden's Paths as a pedagogical tool to students enrolled in his Curriculum Development course. These students consisted of pre-service teachers looking to meet the requirements for certification in the state of Texas. His students' major concerns fell into three categories – usability aspects, limitations of the current data structure (and thus its metaphor), and feature requests. While usability aspects are an important issue to resolve in general, the user concerns mainly focused around better feedback of what the system was doing. The linear nature of Walden's Paths proved to be difficult for the pre-service teachers to adapt to despite the inherently linear nature of traditional classroom instruction. Finally, the students wanted to be able to author their own web pages in the system and embed video from sites like YouTube without having to include advertisements and commentary on the YouTube video. These findings enforced our decision to re-embrace and extend the prototype branching/intersecting paths features from our previous work [4].

2.3 Hypertext Students

The curriculum for the graduate-level course on Hypertext and Hypermedia at Texas A&M University aims to teach students the inception and evolution of hypertext and hypertext systems. The goal of this class is to provide a foundation so that students are able to understand the concepts and research as the field moves forward.

As a preparatory exercise for their final project report, students in the course were asked to create a path based upon the literature surrounding their term projects. This assignment was intended to help the students structure and analyze their related work; provide fellow students with a better base understanding of the literature; and to help their professor gauge the student's progress in researching the related work to their projects.

Unsurprisingly, the resulting paths revealed that students with little to no knowledge of curriculum development or teaching practices created paths that were not ideal knowledge objects to be shared with the greater public for a teaching purpose. These results established a need for vetting good paths. Since the domain of knowledge was focused, this also established a need to socially rate and comment on paths, allowing users with potentially better knowledge of the domain to provide feedback on the paths based on the quality of the material.

3 Subways and Hiking Trails

As the name Walden's Paths implies, the project's original metaphor was one of a path through a woods [12]. The person walking the path could step off and explore "off the beaten path" and then return to the path to continue on. As time went on we began to explore intersecting and branching paths [4]. These structures were then compared to a subway system with the stops being stations that may allow transfer to other lines or exploration of the world just outside the station. In both of these cases these paths were immutable. Paths that were copied and then changed also lost all connection to the original path. However, in real-world hiking trail systems, it is possible to blaze a new trail as a branch, a fork, or an extension. In fact when reading the passage from *Walden* that originally inspired our purposes we see that there is a part of the passage we have ignored, "I had not lived there a week before my feet wore a path from my door to the pondside; and though it is five or six years since I trod it, it is still quite distinct. It is true, I fear, that others may have fallen into it, and so helped to keep it open. The surface of the earth is soft and impressible by the feet of men; and so with the paths which the mind travels." Thoreau was only able to see what was new and unexplored by seeing the contrast between the path he himself had blazed through the forest and the unexplored avenues. These trailblazing activities in real-world hiking are analogous to the path branching, path versioning, and path extension activities. To support our new analogy and issues our informal studies revealed, we redesigned Walden's Path from the ground up.

4 WPv4

Unlike previous Walden's Paths version that featured two monolithic parts – a path creator and a path viewer, WPv4 is an ecology of related interfaces all served by a common thin RESTful API layer. This stands in contrast to many of the new platform as a service solutions (PaaS), backed by Microsoft, through Office Web Apps, and Google, through Google Docs. In Vaquero *et. al'*s working on defining Cloud Computing, they draw a distinction between the Client-Server paradigm of the past and the PaaS paradigm contained in Cloud Computing [13]. However, we believe that even in a PaaS paradigm, a client-server model is still warranted as it help maintains a completely decoupled viewing and creation interfaces from the back end of the system, thus allowing the rapid creation of new interfaces and the possibility to explore ways in which a path can be represented to a user.

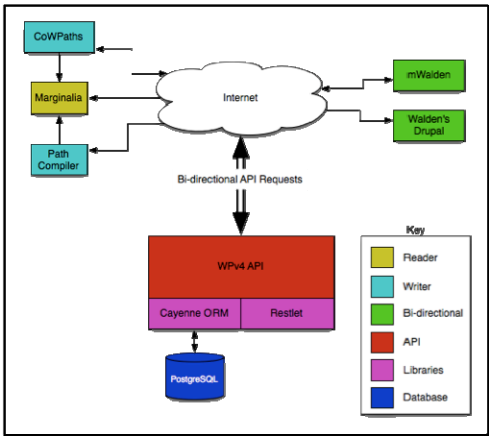


Fig. 1. WPv4 System Diagram

Additionally, the data structures and resulting database were re-designed from the ground up to support a variety of new features. First, group-based access controls allow a user to share paths with a group while restricting general access to the path. For users like the Atmospheric Studies teaching assistants, this allows the group to share paths amongst the group with out requiring a single group login. This is coupled with the second major new feature, a POSIX-like custom permission system. Much like the POSIX octal system of read, write, and execute permissions; we have permissions of read, write, and derive. Permissions can be set for path owners, for zero or more groups, and for the world. While the read and write paths are self-explanatory, the derive permission requires some explanation. As stated before in our discussion of the hiking trails metaphor, a hiker can extend, branch, and fork paths. While extension of paths is handled by the write permission, branching and forking are controlled by derivations. Third, we now support multiple levels of external visibility. Previously, Walden's Paths only supported a published/unpublished notion of visibility. All paths were publically available, but some were unpublished. This was an issue of concern for us as for some purposes, such as an instructor building a reference implementation of a path to compare to student submissions, a path serving as an instructor's key to a path containing question and a personal path that a user may not want to be accessible for security or privacy concerns, a true notion of a private path was needed. However, we did not want to forego a public, unpublished path, as in some use cases, such as in the Atmospheric Studies sections, publishing the paths could lead to reader confusion. Instead we opted to allow different levels of visibility – published, public, and private. Fourth, we support automatic and manual curation abilities, due to our involvement with the Ensemble project, we now support the harvesting of paths over OAI-PMH. However this necessitates a standard of quality to meet guidelines of the NSDL. In order to do this, we allow paths to be explicitly set as vetted by an administrator or ranked as a top path based on social ratings. These ratings are part of our final new feature, support for the standard commenting, rating and tagging social computing interactions.

4.1 Interfaces

By making the core of WPv4 an API with no pre-defined interface, we gain the ability to rapidly experiment with new interface ideas. Currently we are working on five new interfaces, the new default path viewer – Marginalia, a collaborative authoring tool – CoWPaths, a Drupal module for authoring and presentation – Walden's Drupal, a replacement offline path packaging tool – PathCompiler v2, and a mobile path

authoring and view tool – mWalden. The relationship between the path repository, the API and the interfaces is shown in Figure 1. The remainder of this section will discuss each of these interfaces in greater detail.

4.2 Marginalia

In previous versions of Walden's Paths presentation was performed by a monolithic system called the Path Server. The Path Server was responsible for the entire process of retrieving a path from the database then injecting it into the JavaScript sent to the client web-browser to display the path. Marginalia serves as a direct successor to the Path Server and is even compatible as a replacement viewer for WPv3. Unlike the PathServer, Marginalia is a thin JavaScript interface that sends an asynchronous request to a WPv4 server or to a WPv3 PathServer and then renders the path as the request returns. Figure 2 shows Marginalia's presentation of annotations on the left and the page on the right.

Marginalia's name is a reflection of the primary interface difference between it and the PathServer. Marginalia supports pages having multiple annotations rather than being limited to a single annotation. Marginalia chooses to render these annotations as marginalia next to the page in the path rather than above the page as in prior versions of Walden's Paths.

Marginalia is focused on providing a compact and clean interface that is less obtrusive than prior designs while improving annotation readability and intuitiveness of navigation.

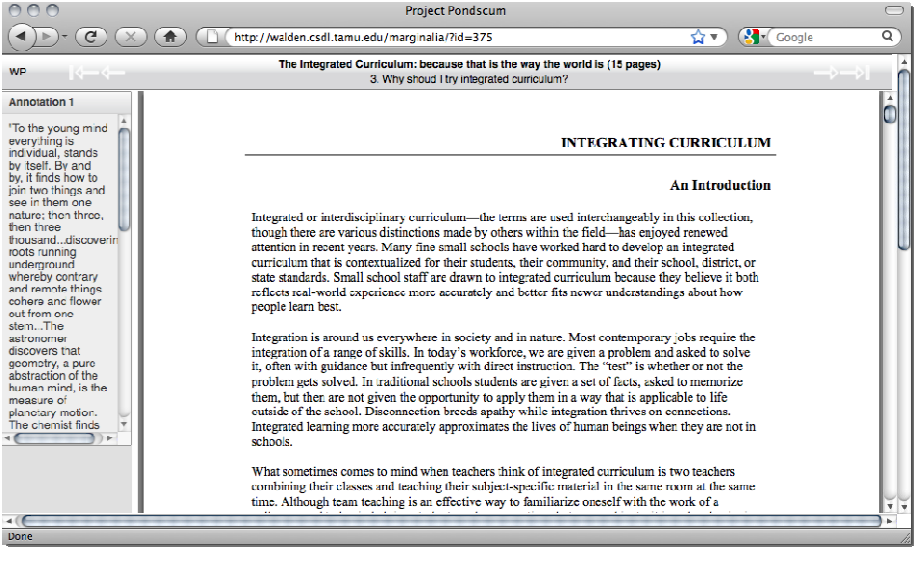


Fig. 2. Marginalia Path Viewing Interface

4.3 CoWPaths

Building upon the desire to share authoring responsibilities, as expressed by the Atmospheric Science users, we set out to create a collaborative authoring tool for Walden's Paths, or CoWPaths. Like our legacy PathPublisher tool, CoWPaths is a web-based authoring system. What distinguishes CoWPaths from its predecessor is the ability to collaboratively author paths, share paths in groups, extend paths, and reuse paths. Users are allowed to create groups, add and remove group members dynamically, and create or add paths they own to groups they belong to. Depending on the desired interactions, permissions within groups can be editing rights to a path or merely read-only access.

Editing a path as a group is done asynchronously with users each having their own authoring interface. Edits to a path by a user are sent to the server asynchronously using WPv4's API and the YUI framework. Once the group finishes the path, they may publish the path publicly, or keep it internal to the group. This also allows groups to be given limited access to the path; for example, teachers can give a group of students read-only access to the path. CoWPaths also gives users an overview of all paths they have access to read. On these paths, users can comment, rate, and tag to help others find applicable paths and judge the quality of them. CoWPaths is the first tool that supports group authoring of paths and is currently exploring collaboration mechanisms and paradigms.

4.4 Walden's Drupal

The Walden's Paths project is part of the NSDL's multi-institute Ensemble pathway project. Ensemble seeks to build tools and provide access to materials to enhance both computing education and computer-based education. For Ensemble, Walden's Paths is both a source of content, in the form of authored paths, and a tool to assemble resources into paths [1].

Since Ensemble is built on Drupal, an Open Source content management system that boasts a highly extensible modular system, the Walden's Paths project decided that in order to provide a consistent user experience and better interoperability with other Ensemble services, that we should design an interface as a Drupal module. This Drupal module allows users to create and publish a Walden's path, and view paths created by others. Our module builds on the Content Construction Kit (CCK) and Views modules, and incorporates jQuery widgets for the interface.

In Drupal, a "Node" is the basic content element. A node can be made into many different kinds of content, such as: a blog entry, a page or a forum topic. In our system, we have two content types: path and page. A path serves as a container for the pages in a path and metadata, such as a title, on the path as a whole. A page content type is an encapsulation of one resource on the Internet, including the name of current resource, the URL of to the resource and annotations added by the node creator.

This module allows users to create and publish a Walden's path to other users, using the same mechanism in Drupal as a publishing a blog entry. Other users then can view published Walden's paths or harvest them over the OAI-PMH modules enabled on the Ensemble Project's Drupal server. Figure 3 shows a page in Walden's Drupal. Finally, via the WPv4 API, paths can be shared between a WPv4 server and an instance of Walden's Drupal.

4.5 PathCompiler v2

Early in the development of Walden's Paths we encountered a situation where users needed to view paths despite slow and unreliable connections. Our solution, PathCompiler, created self-contained paths that enabled users to view paths offline [6].

PathCompiler v2 is a re-implementation of the original PathCompiler, but leveraging the WPv4 tools and targeting the use of Walden's Paths in a more global aspect by allowing a portable standalone path that can be loaded on to whatever device a user wishes to view a path regardless of its connectivity. PathCompiler v2 connects to the API of WPv4 to get paths that have been defined by Walden's Path users. PathCompiler v2 then saves all necessary materials on the local machine along with the path metadata in order to show World-Wide Web pages off-line. Since the modern Internet is not just static pages, but is full of multimedia materials, pictures and scripts, the system must be able to find included materials and a portion of linked to pages off the path to replicate the online Walden's Paths experience.

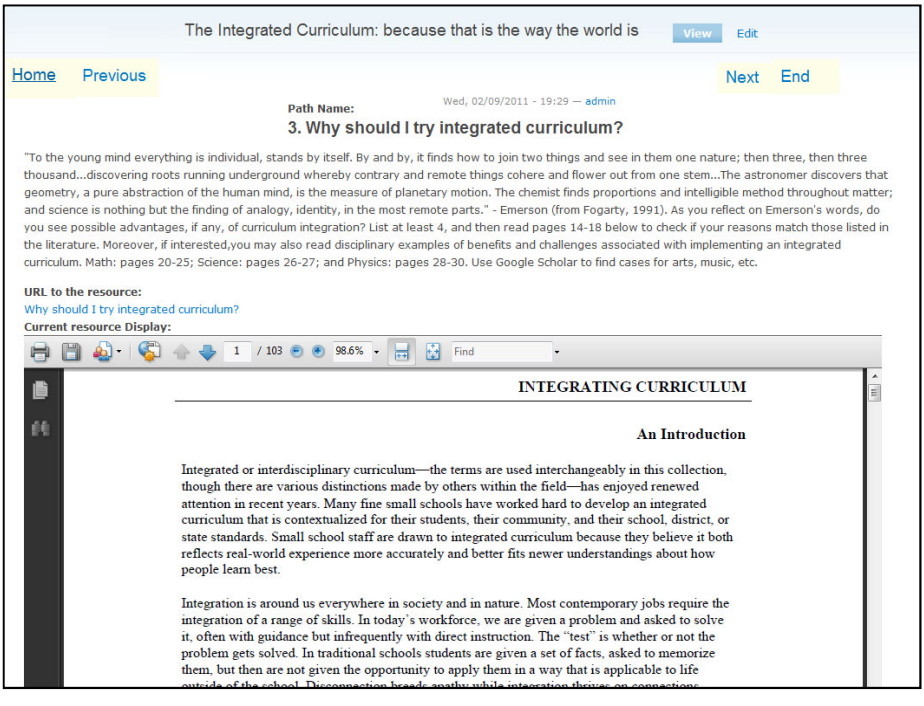


Fig. 3. Path Viewing in Walden's Drupal

PathCompiler v2 is implemented in Python and emits a set of web pages and related materials that can be accessed through a specially modified version of the Marginalia viewer.

PathCompiler v2 is of interest to the Ensemble Project. Ensemble needs to represent collections of materials that are found on the web but then gathered locally to ease long-term use. Additionally, PathCompiler can be coupled with mWalden to

provide a solution for paths on devices with limited connectivity such as the growing base of Internet capable phones both in the developed and developing worlds [8].

4.6 mWalden

Walden's Paths in the past has been primarily a traditional web-based interaction. This web-based interaction works well in a traditional learning environment where a teacher can present a path to the class on a projector or have students in a lab or on their home computer view the path. However, for a growing percentage of the world's Internet users, a mobile device is their primary or only means of Internet access [8]. This emerging sector is even more important in developing countries where Internet access, even on mobile devices, is limited by speed and availability [5]. This new era of increasing ubiquity of mobile devices as a primary means of access to information further supports the need for the Client-Server model since native applications, like mWalden, reduces the load placed on the network by an application while providing a richer user experience.

mWalden provides an Android-based application that can author paths, view paths stored on a WPv4 server, or emitted from PathCompiler v2. This allows the creation of guided narrative on top of web pages that is better suited for the mobile environment.

Additionally, the GPS capability of many cellular devices can be harnessed to provide a new kind of path – a geo-located path. In these paths each node is tied to real-world coordinates and can be used to explore physical areas with supplemental materials and guided narration provided to the user. These paths can be explored in situ as a walking tour or exploration and they can be explored ex situ using a map-based interface to maintain the linkage between place, web content and annotation.

5 Conclusions and Future Work

As we stated at the beginning of this paper, our evaluations illuminated a set of issues regarding Walden's Paths. To review these issues can be roughly described as regarding authorship, presentation, portability and indexabilty. In order to approve authorship particular in collaborative work scenarios, we created CoWPaths, out collaborative authoring tool for paths. The Marginalia web-interface's support for multiple annotations and Web 2.0 resources, such as videos and PDFs, address issues of presentation in traditional PC-based path consumption scenarios. Additionally, Walden's Drupal is addressing issues or indexability, presentation, and authorship for the existing Drupal-based websites in addition to the special case of the Ensemble Project's computer science education pathway. We are investigating offline presentation and techniques to devise notions of limited freedom to roam in PathCompiler v2. Finally, mWalden seeks to bring Walden's Paths to mobile devices and the emerging communities that depend on small devices for information access. Work on mWalden, along with the accompanying issues of authorship and presentation, is exploring novel forms of path-based interactions enabled by the capabilities of mobile devices.

In the future, we intend to expand the choices of interfaces that users have at their disposal. For CoWPaths, we are planning to experiment further in to collaborative authoring of pedagogically sound paths. Simultaneously, we will continue to use mWalden to build paths that enhance exploration of physical spaces.

In conclusion, WPv4 and its interfaces are providing a platform for development of new user interactions with paths. While this client-server design may seem to violate the tenets of the current trend of platform as a service, instead we have shown that a decouple and distributed platform as a service has improved our ability to experiment in areas that we had not explored in the past such as collaborative authoring, mobile interfaces, and location-aware paths.

Acknowledgements. This material is based upon work supported in part by the National Science Foundation under Grant No. DUE 08-40713.

References

[1] Brusilovsky, P., Cassel, L., Delcambre, L., Fox, E., Furuta, R., Garcia, D.D., Shipman III, F.M., Bogen, P., Yudelson, M.: Enhancing digital libraries with social navigation: The case of ensemble. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 116–123. Springer, Heidelberg (2010)

[2] Bush, V.: As We Think. Atlantic Monthly, Boston (1945)

[3] Dave, P., Bogen, P., Karadkar, U., Francisco-Revilla, L., Furuta, R., Shipman, F.: Dynamically growing hypertext collections. In: HYPERTEXT 2004: Proc. of the 15th ACM Conf. on Hypertext and Hypermedia. ACM, Santa Cruz (2004)

[4] Dave, P., Karadkar, U., Furuta, R., Francisco-Revilla, L., Shipman, F., Dash, S., Dalal, Z.: Browsing intricately interconnected paths. In: Proc. of the 14th ACM Conf. on Hypertext and Hypermedia. ACM, Nottingham (2003)

[5] Donner, J., Gitau, S.: New paths: exploring mobile-centric internet use in South Africa. In: Int. Comm. Assoc. Conf., Chicago, IL (2009)

[6] Furuta, R., Shipman, F.I., Marshall, C., Brenner, D., Hsieh, H.: Hypertext paths and the World-Wide Web: experiences with Walden's Paths. In: Proc. of the 8th ACM Conf. on Hypertext, ACM, Southampton (1997)

[7] Kim, S., Slater, M., Whitehead, J.: WebDAV-based Hypertext Annotation and Trail System. In: Proc. of the 15th ACM Conf. on Hypertext and Hypermedia, ACM, Santa Cruz (2004)

[8] Kreutzer, T.: Generation mobile: online and digital media usage on mobile phones among low-income urban youth in South Africa. U. of Cape Town, SA (2009)

[9] Moody, P.: WebPath: Sharable Personalized Guided Web Tours. IBM Watson Research Center, Cambridge, MA, Lotus Technical Report, no: 98-09

[10] Schraefel, M., Smith, D., Owens, A., Russell, A., Harris, C., Wilson, M.: The Evolving mSpace Platform: Leveraging the Semantic Web on the Trail of the Memex. In: Proc. of the 16th ACM Conf on Hypertext and Hypermedia. ACM, Salzburg (2005)

[11] Shipman, F., Furuta, R., Brenner, D., Chung, C., Hsieh, H.: Guided paths through web-based collections: design, experiences, and adaptations. J. Am. Soc. Inf. Sci. 51(3)

[12] Shipman, F., Furuta, R., Marshall, C.: Generating Web-based presentations in spatial hypertext. In: Proc. of the 2nd Int. Conf. on Intelligent User Interfaces. ACM, Orlando (1997)

[13] Vaquero, L., Rodero-Merino, L., Caceres, J., Lindner, M.: A break in the clouds: towards a cloud definition. In: SIGCOMM Computer Comm. Review. ACM, New York (2008)

Understanding the Dynamic Scholarly Research Needs and Behavior as Applied to Social Reference Management

Hamed Alhoori and Richard Furuta

Center for the Study of Digital Libraries and
Department of Computer Science and Engineering
Texas A&M University, USA
{alhoori,furuta}@tamu.edu

Abstract. We conducted a study with an objective to learn more about the dynamic information needs, information-seeking behavior, information use and other scholarly activities of researchers. Our focus was on the collaborative and social usage and on the social reference managers. We compared the current practices and strategies of scholars and researchers from multidisciplinary research areas. Our findings provide valuable insights and augment the understanding of how the social web is having a significant effect on the current researchers' activities and digital libraries.

Keywords: Scholarly communication, Research community, Digital Libraries, Information Seeking, Social web, Social reference management, Scholarly references, Social Bookmarking, Knowledge management, Literature review.

1 Introduction

Living and working in a dynamic knowledge society requires deep awareness and access to the best contents and real-time research results to assist and improve strategic decision-making. Understanding the dynamic scholarly activities and their related pathways taken by researchers plays a significant role in developing library collections and services. We have studied several scholarly activities such as: searching, organizing and retrieving articles or notes, collaboration among researchers, usage of social reference managers (SRMs) and its effect on the scholarly activities.

However, while the success of digital libraries increases the materials available to the scholar it also increases the complexity of the scholarly research environment. In this, locally-generated digital libraries serve as a reflection of the frequently-noted global explosion of information—more than 25,000 peer-reviewed research journals exist worldwide, across all disciplines and languages, publishing about 2.5 million articles per year [1]. Filtering and discovering the best results in a short time can be like finding a needle in a haystack.

Traditional libraries provide services to their users while prohibiting them from contributing, and most early digital libraries adopted this viewpoint as well. This results in a considerable loss of external knowledge. However, the current state of the

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 169–178, 2011.

art is moving toward two ways of interaction, where users both can benefit from the available knowledge and also can contribute to it. Digital library contents moved from being accessed by isolated databases, to a more social and collaborative environments. Instead of being limited by only storing content, personal copies and notes in a personal computer or server, researchers are moving to share and annotate links to their favorite research content on the cloud. Social bookmarking sites [2] have gained visibility in the past few years with these sites reaching thousands of users and containing millions of bookmarks and tags. Beyond increases in effectiveness in finding resources, social bookmarking systems help users to become aware of more relevant information that is available.

Editors and reviewers openly lament the inadequacy of literature reviews in manuscripts submitted for journal publication [3]. Conducting comprehensive research in this era requires having the tools that support the researcher to know the related papers and discover new content. Social bookmarking for researchers [4] or the so-called SRM communities (e.g., citeulike [5] and Mendeley [6]) are playing a significant role in the conduct of research. Retrieving the best results by searching and browsing is no longer the best way to find relevant information.

Algorithms are being designed and developed to suit the scholarly and social community needs such as filtering and discovering items of interest [7], connecting with like-minded researchers and getting recommendations based on their digital libraries content and related work [8]. Other efforts have been directed to reduce the web spam and redundancy [9] that started targeting more specific communities, such as the scholarly world, and introduced a variety of features to fight spam in social bookmarking systems. We have investigated the precision outcomes of a hybrid bibliography system created by an online digital community to support the creation of scholarly bibliographies [10]. Our experimental results indicate that using online reputation based social collaboration improve the quantity and usage of scholarly bibliography and improve the quality and creditability of social citations sites.

This paper is structured as follows. We discuss the related work in Section 2. Section 3 explains the methodology we used. We present and discuss our findings and results in section 4. In section 5 we conclude and highlight some of the future work.

2 Related Work

Abundant studies have been conducted in various disciplines such as geoscience [11], chemistry [12], agricultural and biological sciences [13], medicine and health sciences [14], public health [15], veterinary medicine [16] , law [17] humanities [18] to better understand the dynamic information needs, information-seeking behavior, information use and other scholarly activities of researchers, scientists[19], engineers [20], academic scholars [21], undergraduate students [22], graduate students [23].Several methods were used to examine the scholarly activities using quantitative studies (e.g. surveys [24]), qualitative studies (e.g. interviews) [25], ethnographic observational studies [26] or combination. In [27] they used transactional log studies. Brown used a combination of the e-mail survey and content analysis methods [23]. Some studies used citation analysis [28].

Brown [29] investigated the differences and similarities in information-seeking behavior of academic scientists in four disciplines: astronomy, chemistry-biochemistry, mathematics, and physics. All scientists responding scanned the latest issues of journals to keep abreast of current developments in their fields. The mathematicians surveyed indicated an additional reliance on monographs, preprints, and attendance at conferences and personal communication to support their research activities. Hallmark [30] described the methods of access and retrieval of recent journal articles cited by geoscientists and chemists who work in academia, government, and industry. The study found that the majority of scientists had developed effective new patterns of searching for useful references.

In [31] a multi-disciplinary study explored graduate students' information behavior related to their process of inquiry and scholarly activities. They found that their skills and decisions are influenced directly by professors, other students, librarians and Internet usage. Other results were that the lack of sophistication in finding and using resources and course requirements affect students' information behavior and that findings vary across disciplines and between programs. In addition, some graduate students mentioned influences such as difficulty locating information or the need for convenience and speed.

A number of studies showed that researches are not aware of or familiar with some of the resources, services and electronic search tools available for them through the library and generally do not consult librarians regarding their information needs [13][32]. Part of the difficulties encountered by researchers in using resources appear to stem mainly from a lack of training [33] .The findings in [32] indicate that guidance in the use of library resources and services is necessary to help students meet some of their information requirements. The study found that journals, library books, and textbooks are the most popular sources of information for course work and research, and that students need to be taught how to use available library resources and services.

Hoffmann, et al. [34], found that graduate students wanted to learn about strategies for finding information, bibliographic management tools, and tools for keeping current with scholarly literature. Students preferred online instruction, although in-person workshops were also found to be valuable. Workshops have been held to support researchers' activities such as using particular tools [35] (e.g., bibliographic management software). The authors of [36] created literature review workshops to serve graduate students from a wide range of subject disciplines at a point of shared need. They identified some of the gaps graduate students have in their knowledge about library services. Nicholas, et al. [37] compared the information seeking behavior of the users' logs of four universities using the OhioLINK journal system and found large differences especially between the research and teaching universities.

Niu and Hemminger, et al. [38], surveyed 2,063 academic researchers in natural science, engineering, and medical science from five research universities in the United States to understand different aspects of researchers' information-seeking behavior. Descriptive statistics were reported by institutions to compare differences among universities. Findings reflected the dominant utilization of electronic methods

for searching and accessing scholarly content. Differences in information-seeking behavior amongst universities were not as clear as those amongst disciplines and demographics. A notable trend is that novel forms of scholarly communication such as collaborative information sharing technology are evolving gradually. They expect that this may be the beginning of a more significant transformative change, particularly in sharing information within laboratories or groups or amongst multisite collaborations. Many professors have begun utilizing blogs, wikis and multimedia to communicate with their colleagues or students. Collaborative search systems, academic social bookmarking systems, open shared rankings and reviews, open access journals, and online sharing bibliographic databases and annotations were all examples of new scholarly communication information technologies. The adoption of these was consistent among the respondents across the five universities.

Most studies were limited on a single campus and did not consider the dynamic changes in scholars information needs and behavior, opportunities and challenges of the social web, or were before its emergence; there were no, or limited, ways of sharing, collaboration, connecting researchers, discovering and recommending content. We have investigated how changes in the technologies available to research communities addressing the use of social media can be used to the benefit of researchers, supporting their overall research progress and outcome. Our research questions included:

- How do researchers search, select, and manage their information sources?
- What difficulties researchers are facing during literature review process?
- How SRM influenced the literature review process?
- What are the current scholarly research needs?

3 Methodology

Our study used two methods of data collection: a qualitative research method using interviews and a quantitative method using an online survey. The same set of questions was used as basis for both methods. Before the methods were carried out, seven researchers reviewed the questions dataset before in order to assess its effectiveness and completion time required. Minimal modifications were made based on this feedback. Participation in both studies was confidential and voluntary. Participants were able to withdraw at any time.

We compared the similarities and differences of researchers considering their scholarly activities. In our interviews, a set of eight randomly selected faculty members from different disciplines on campus (see Table1) were invited to participate in personal interviews. Interviews lasted 45-60 minutes. Most of the faculty interviewed, supervised a research group with active researchers. The interviews started by discussing the current practices in the research group using open-ended questions. Then we moved to cover the unanswered questions from our set of questions. Interview sessions were manually transcribed. Transcriptions of all the discussions were manually coded.

Table 1. Scholars IDs and rank

ID	Research Area	Rank
1	Statistics	Associate Professor
2	Petroleum engineering	Assistant Professor
3	Biology	Assistant Professor
4	Management	Assistant Professor
5	Chemical Engineering	Professor
6	Microbiology	Professor
7	Education	Associate Professor
8	Computer Science	Assistant Professor

The survey was sent to different university departments and social reference manager groups. In the survey, samples were random, independent and quite large, so we used statistical hypothesis testing techniques to investigate how using the SRMs affects the research process. We used Pearson's chi-square test (X^2) and Analysis of variance (ANOVA).

4 Results and Discussion

4.1 Interview

Scholars differ in their reading habits, but in general they agreed that they skim the paper first by reading its abstract, conclusion or results section, and then decide if they will read the complete paper. While some get lost while moving between different papers and references, some kept notes and focus on high impact papers. They agreed that they stop working on literature review when they have enough information and literature content starts to repeat itself.

Most scholars said that they come across at least few articles that would add value to their completed or published work if they knew it exist. In line with conclusions that previous studies showed researchers are finding some difficulties locating their needs as P1 commented:

> "I know the information is there, but I don't know how to reach it in a short period of time".

Scholars mentioned a number of difficulties during the literature review process. Several scholars complained about the repeated results during the continuous search process as P2 stated:

> "I would like to have a way to remove the previously viewed results from my new search results or when checking for new citations. Worse than that, when I get some search results that are already stored in my articles collection or reference manager and I start to view them again since my collection is huge and I can't remember all articles, which is totally useless".

In saving and organizing articles some scholars were still printing articles, and when asked why they didn't move to use advance ways, they said they have been using it for long time and don't want to jump between several tools as P3 commented:

> "I print all the papers I need and organize them using authors' names. Although it may take some time to find what I need, however this way works for me since my graduate school".

Few scholars feel satisfied with organizing their papers and notes using folders and text files, as P5 explained:

> "I have been using folders to organize my papers and notes based on projects. I know all my folders and when I need anything, I can go back to the project and to the subfolders".

One scholar was even using general organizing tools as P6 explained:

> "I am happy using my old file organizing tool version 1.0."

Several scholars used reference managers and share references among their groups. However, others when asked why they don't use a reference manager most were concerned with the learning curve time and possibility of delaying their work as P6 commented:

> "I have used the free reference manger provided by the university library. Although it was good, it needs a license and continuous update which delayed my work especially when I move between several places."

Reference managers are becoming an integral tool during the research progress. When asked why P4 was using a reference manager he explained:

> "I have around 12,000 articles and I am daily adding few more. I also share some with other scholars".

Scholars take notes on their printed articles or in reference managers. Others were using some online note-taking sites or emails. Few were even using text files and attaching to them all saved articles, notes or ideas. When asked how she remembers where a paper or saved note is, P1 said:

> "I have a strong memory, so I know most of my printed papers and the attached notes".

To keep up to date some researchers do a repeat manual search and were not aware of alerts as P5 stated:

> "I repeat some searches from time to time and check if there are any new articles to read. Having a tool that can provide me with my research interests can save me a lot of time".

Most scholars have collaborated with other scholars. A major reason was to expand their knowledge and speed the work progress. They select whom they want to collaborate based on others' reliability and ability to collaborate. Some scholars didn't know how the SRM works, and they didn't want to spend time exploring them as P3 commented:

> "I am busy with my work and getting my tenure. I don't want to spend time using SRM and adding friends so that I can get the articles recommendations".

However, after knowing how easy SRM works, some were willing to test them and later sends us a thank you letter. A number of researchers expressed regret about their lack of awareness regarding SRMs. P7 was surprised to hear about the SRMs:

> "I never heard about the SRM. Actually, I have searched Facebook applications to share my references online with friends, but end up using emails and Google docs".

SRM users showed some concerns about accuracy of the bibliographic data as P8 explained:

> "I usually found some errors, missing bibliographic data or duplicate social bookmarks. So, I usually verify its data with the article published press website".

4.2 Survey

156 researchers participated in the online survey as follow (17 faculty members, 5 postdoctoral, 84 doctoral students, 28 master students, 22 undergraduate students). There were 124 male respondents and 32 female; 64% were between 26 and 34 years old. Participants were from 13 different disciplines. We applied several tests to find any significance in the results (p=p-value). We compared how researchers saving methods influence other scholarly activities. Saving methods were using computer folders/directories, reference managers or SRM.We found that SRM users differ significantly from other users in how they search for articles (X^2=44.31, df=4, $p < 0.001$). While most researchers used general or specific search engines, 40% of the SRM users search within SRM environment. They explained using SRM to search since it has more relevant and newer results, connecting with like-minded researchers or even accurate bibliographic data. SRM users also use tags more often than other users. We found a significant relationship between using SRM and tags usage (X^2=19.032, df=1, $p < 0.001$). SRM users were able to find more related articles to their research interests than other users. However, there was no significant relationship between using SRM and finding related topics (X^2=2.11, df=1, $p < 0.05$).

Publications overloading is still a major challenge for most researchers (78%) even for SRM users. However, there was no significant relationship between publications overloading and saving methods ($X^2 = 0.79$, df=2, $p < 0.05$) or between publications overloading and how users organize their articles (X^2=1.35, df=1, $p < 0.05$). Organizing methods were divided into folders, tags or visual tools. Some SRM users show an interest using visual tools but there were no strong evidence of a relationship.

Researchers who use folders get lost more often when reading and moving between articles. We found a significant relationship between saving methods and getting lost while reading and navigating between articles (X^2=12.71, df=6, $p < 0.05$). We found another significant relationship between saving methods and taking notes on printed papers (X^2=5.64, df=1, $p < 0.05$). Researchers that take notes on papers constituted 68% of those that use folders, 50% of those that use reference managers and only 19% of those that use SRM. Furthermore, we found a significant relationship between using SRM and taking notes within the SRM (X^2 = 17.03, df=1, $p < 0.001$).

We found a significant relationship between saving methods and researchers first way to retrieve articles (search or browse) they read recently (X^2=9.98, df=2, $p < 0.05$). Those that start with search were only 31% of those that use folders, 50% of those that use reference managers and 63% of those that use SRM. There was a significant relationship between saving methods and whether they collaborate with other users or not (X^2=6.82, df=2, $p < 0.05$). Those that collaborate were 59% of those that use folders, 80% of those that use reference managers and 81% of those that use SRM.

Most researchers collaborate with others (67%) with different reasons mentioned: share and expand knowledge, make new connections, increase possibility of getting funds, motivation, speedup the work or publish more. Researchers that don't collaborate provide different reasons such as, busy with their research, hard to compile/ synchronize the work or don't know other users with similar interests.

Finally, we found a strong evidence that saving methods have an effect on researchers satisfaction while searching ($F=37.80$, $P < 0.001$), while retrieving articles ($F=4.67$, $P < 0.05$) and organizing articles ($F=4.66$, $P < 0.05$).

5 Conclusion and Future Work

This study investigated the current practices and dynamic scholarly activities. It illustrates the remarkable effect that SRMs have had on the scholarly process. The SRM plays a significant role in finding and organizing scholarly articles, connecting researchers, improving collaboration, providing article recommendations, increasing scholarly awareness and revolutionizing scientific communication. SRMs have the opportunity to meet more researchers' information needs and improve their information-seeking behavior.

Academic libraries need to increase the awareness of research technologies available, especially since SRMs are relatively new tools; most users get familiar with a tool and need to be motivated to change to better technologies later. SRM allows the research community to gain many benefits and could have enormous impact in the future on the overall research process.

A 2006 study [31] found that nearly all graduate students (96%) reported that academic staff (e.g., advisers, professors and committee members) influence their research and information seeking. We would like to investigate if SRM has any significant effect on research groups toward building online collaborative research communities. We intend to investigate more the effects of SRMs on the research process and develop a collaborative research model of dynamic strategies. We plan also to investigate how visual tools can influence SRM usage.

References

1. Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C., Hilf, E.: The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. Serials Review 34, 36–40 (2008)
2. Hammond, T., Hannay, T., Lund, B., Scott, J.: Social Bookmarking Tools (I). DLib Magazine 11 (2005)
3. Boote, D.N., Beile, P.: Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. Educational Researcher 34, 3–15 (2005)
4. Farooq, U., Song, Y., Carroll, J.M., Giles, C.L.: Social Bookmarking for Scholarly Digital Libraries. IEEE Internet Computing 11, 29–35 (2007)
5. Emamy, K., Cameron, R.: CiteULike: A Researcher's Social Bookmarking Service. Ariadne 51 (2007)
6. Henning, V., Reichelt, J.: Mendeley - A Last.fm For Research? In: 2008 IEEE Fourth International Conference on eScience, pp. 327–328. IEEE, Los Alamitos (2008)
7. Bogers, T., Van Den Bosch, A.: Recommending scientific articles using citeulike. In: RecSys 2008: Proceedings of the 2008 ACM Conference on Recommender Systems, pp. 287–290. ACM, New York (2008)
8. Dicheva, D., Dichev, C.: Finding Resources and Collaborators within Digital Collections. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 631–638 (2010)
9. Heymann, P., Koutrika, G., Molina, H.G.: Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. IEEE Internet Computing 11, 36–45 (2007)
10. Alhoori, H., Alvarez, O., Furuta, R., Muñiz, M., Urbina, E.: Supporting the creation of scholarly bibliographies by communities through online reputation based social collaboration. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 180–191. Springer, Heidelberg (2009)
11. Bichteler, J., Ward, D.: Information-seeking behaviour of geoscientists. Special Libraries 80, 169–178 (1989)
12. Davis, P.M.: Information-seeking behavior of chemists: A transaction log analysis of referral URLs. Journal of the American Society for Information Science and Technology 55, 326–332 (2004)
13. Kuruppu, P.U., Gruber, A.M.: Understanding the Information Needs of Academic Scholars in Agricultural and Biological Sciences. The Journal of Academic Librarianship 32, 609–623 (2006)
14. Davies, K.: The information-seeking behaviour of doctors: a review of the evidence. Health Information and Libraries Journal 24, 78–94 (2007)
15. Revere, D., Turner, A.M., Madhavan, A., Rambo, N., Bugni, P.F., Kimball, A., Fuller, S.S.: Understanding the information needs of public health practitioners: a literature review to inform design of an interactive digital knowledge management system. Journal of Biomedical Informatics 40, 410–421 (2007)
16. Pelzer, N.L., Wiese, W.H., Leysen, J.M.: Library use and information-seeking behavior of veterinary medical students revisited in the electronic environment. Bulletin of the Medical Library Association 86, 346–355 (1998)
17. Makri, S.: Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. Information Processing & Management 44, 613–634 (2008)
18. Barrett, A.: The Information-Seeking Habits of Graduate Student Researchers in the Humanities. The Journal of Academic Librarianship 31, 324–331 (2005)
19. Tenopir, C., King, D.W., Boyce, P., Grayson, M., Zhang, Y., Ebuen, M.: Patterns of Journal Use by Scientists through Three Evolutionary Phases. DLib Magazine 9 (2003)
20. Hertzum, M., Pejtersen, A.M.: The information-seeking practices of engineers: searching for documents as well as for people. Information Processing & Management 36, 761–778 (2000)

21. Hemminger, B.M., Lu, D., Vaughan, K.T.L., Adams, S.J.: Information seeking behavior of academic scientists. Journal of the American Society for Information Science and Technology 58, 2205–2225 (2007)
22. Warwick, C., Rimmer, J., Blandford, A., Gow, J., Buchanan, G.: Cognitive economy and satisficing in information seeking: A longitudinal study of undergraduate information behavior. Journal of the American Society for Information Science and Technology 60, 2402–2415 (2009)
23. Brown, C.: Where Do Molecular Biology Graduate Students Find Information? Science Technology Libraries 25, 89–104 (2005)
24. Hirsh, S., Dinkelacker, J.: Seeking information in order to produce information: An empirical study at Hewlett Packard Labs. Journal of the American Society for Information Science and Technology 55, 807–817 (2004)
25. Bolduc, A.: Surveying user needs in an international context: A qualitative case study from the ILO, Geneva. The International Information Library Review 40, 1–9 (2008)
26. Haglund, L., Olsson, P.: The Impact on University Libraries of Changes in Information Behavior Among Academic Researchers: A Multiple Case Study. The Journal of Academic Librarianship 34, 52–59 (2008)
27. Nicholas, D., Huntington, P., Jamali, H.R., Watkinson, A.: The information seeking behaviour of the users of digital scholarly journals. Information Processing & Management 42, 1345–1365 (2006)
28. Smith, E.T.: Assessing collection usefulness: An investigation of library ownership of the Resources Graduate Students Use. College & Research Libraries 64, 344–355 (2003)
29. Brown, C.M.: Information seeking behavior of scientists in the electronic information age: Astronomers, chemists, mathematicians, and physicists. Journal of the American Society for Information Science 50, 929–943 (1999)
30. Hallmark, J.: Access and Retrieval of Recent Journal Articles: A Comparative Study of Chemists and Geoscientists. Issues in Science and Technology Librarianship 40 (2004)
31. George, C., Bright, A., Hurlbert, T., Linke, E.C., St. Clair, G., Stein, J.: Scholarly use of information: graduate students' information seeking behaviour. Information Research 11, 1–19 (2006)
32. Fidzani, B.T.: Information needs and information-seeking behaviour of graduate students at the University of Botswana. Library Review 47, 329–340 (1998)
33. Vibert, N., Rouet, J., Ros, C., Ramond, M., Deshoullieres, B.: The use of online electronic information resources in scientific research: The case of neuroscience. Library & Information Science Research 29, 508–532 (2007)
34. Hoffmann, K., Antwi-Nsiah, F., Feng, V., Stanley, M.: Library Research Skills: A Needs Assessment for Graduate Student Workshops. Issues in Science and Technology Librarianship 53 (2008)
35. Harrison, M., Summerton, S., Peters, K.: EndNote training for academic staff and students: the experience of the Manchester Metropolitan University Library. New Review of Academic Librarianship 11, 31–40 (2005)
36. Rempel, H.G., Davidson, J.: Providing Information Literacy Instruction to Graduate Students through Literature Review Workshops. Issues in Science and Technology Librarianship 53 (2008)
37. Nicholas, D., Huntington, P., Jamali, H.R.: Diversity in the Information Seeking Behaviour of the Virtual Scholar: Institutional Comparisons. The Journal of Academic Librarianship 33, 629–638 (2007)
38. Niu, X., Hemminger, B.M., Lown, C., Adams, S., Brown, C., Level, A., McLure, M., Powers, A., Tennant, M.R., Cataldo, T.: National study of information seeking behavior of academic researchers in the United States. Journal of the American Society for Information Science and Technology 61, 869–890 (2010)

Experiment and Analysis Services in a Fingerprint Digital Library for Collaborative Research

Sung Hee Park[1], Jonathan P. Leidig[1], Lin Tzy Li[1,3,4], Edward A. Fox[1], Nathan J. Short[2], Kevin E. Hoyle[2], A. Lynn Abbott[2], and Michael S. Hsiao[2]

[1] Digital Library Research Laboratory, Department of Computer Science
[2] Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA 24061, USA
[3] Institute of Computing, University of Campinas, Campinas, SP, Brazil, 13083-852
[4] CPqD Foundation, Campinas, SP, Brazil, 13086-902
{shpark,leidig,lintzyli,fox,nshort21,kevin87,abbott,mhsiao}@vt.edu

Abstract. Fingerprint management systems support millions of images and complicated but imperfect image identification algorithms. The forensic community requires a set of digital library services to support large image collections, execute identification algorithms, and analyze experiments that test identification algorithms in development. We present a model and prototype system capable of testing and analyzing fingerprinting algorithms in terms of identification performance based on matches of a known image to partial images, distortions of the images, and sub-regions of the images. These services are provided based on our framework for composing a set of services and a fingerprint image collection. The prototype will be useful in collaborations connecting several algorithm development efforts, and in composing an experimentation workflow. We also describe extensions of these services into other domains.

Keywords: fingerprint collections, algorithms, experiments, analyses.

1 Introduction

Fingerprint identification has been a staple of forensics and criminal justice for more than a century. National fingerprint collections contain millions of images and are searched by extracting details on the unique ridge structure and minutiae. Minutiae are common features in a fingerprint. Developing algorithms for image retrieval, comparison, identification, compression, and analysis remains an active field. Fingerprint management systems have been developed to maintain recorded prints in civilian, military, and criminal collections in addition to crime scene collections. The FBI fingerprint system contains around 66 million criminal and 25 million civil images [4]. The largest fingerprint collections are proprietary or government-owned and are not released for public usage. Researchers currently lack a digital library (DL) for human training and the developing, testing,

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 179–191, 2011.

and training of fingerprint identification algorithms. Such a DL would be useful for studying and modifying existing algorithms that detect minutiae, ridges, sub-images, and full images. To support identification tasks, digital library services are required to expose large image collections, support algorithms, and maintain algorithm analysis experiments.

We propose a fingerprint digital library with services to manage collections, algorithms, analysis experiments, and experiment results as shown in Fig. 1. The provided services allow researchers to gauge the quality of matching algorithms for distorted and sub-region images. While the services are tailored to fingerprint images and specific algorithms in the prototype implementation, we propose a generic model for maintaining scientific data and plugging tools or algorithms into a testing environment. *The goal of this work is to present a model and prototype of an end-to-end image-based DL experimentation and analysis service.*

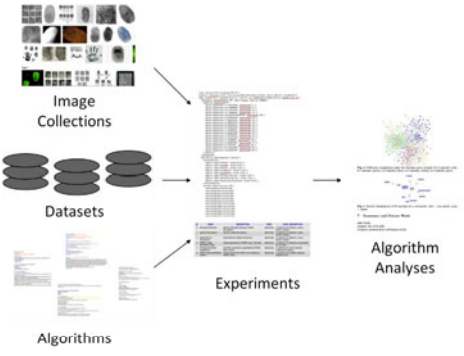


Fig. 1. Workflow framework: fingerprint images, feature or distortion datasets, and image-processing algorithms are used in experiments to analyze algorithm performance

This paper describes fingerprint images and collections in Section 2, algorithm and analysis services in Section 3, a proposed framework and prototype experiment DL in Section 4, and then a summary and plan for future work.

2 Fingerprint Image Collections

Current research into fingerprinting algorithms is hampered by the lack of a large, well-defined testing dataset and infrastructure for executing new or revised algorithms. Testing ridge mapping algorithms for quality assurance is currently a human-intensive endeavor due to computational errors because of fuzziness and blurriness in distorted images. On a similar note, ascertaining the quality of feature extraction algorithms also requires comparisons to human identified features.

2.1 Fingerprint Images

For over a century, fingerprints have been classified by features clear to human observers. Historical classifications of fingerprint features include combinations

of archs, loops, and whorls. Finer-grained features are labeled minutiae, which are several points along fingerprint ridges. Ridge points include terminal ends, islands, and forking locations, called bifurcations, see Fig. 2. Fingerprint subregions can be differently affected by finger pressure, direction of movement, stretches, and humidity, with several levels of quality compared to an original print. Humidity levels include wet, proper, and dry, see Fig. 3.

Images used in fingerprinting are often not perfect specimens. In a forensic setting, crime scene evidence might be a partial, smeared, distorted, or overly humid fingerprint. In order to train algorithms and software with poor quality field images, distorted example images must be generated to analyze each potentially troublesome factor. Several types of distortions may transform high-quality fingerprint images on record to what are found as poor-quality images in actual evidence. A distorted image may have displacements in the x and y directions on a plane, a distortion common with partial images. Rotations and skin plasticity also distort images, as is shown in Fig. 4. Notice that when the skin's focal point of contact is where pressure is applied, rotations and plasticity will cause multiple portions of an image to be skewed differently.

In this study, we use real fingerprint images from Fingerprint Verification Competition (FVC) 2000 DB1 and DB2, consisting of 880 images (110 different individuals who are 20 to 30 year-old students, about 50% male). For latent image analysis, we used the National Institute of Standards and Technology (NIST) Special Database 27 as a real dataset of fingerprint minutiae from latent and matching tenprints. These datasets are widely used by researchers and trainers.

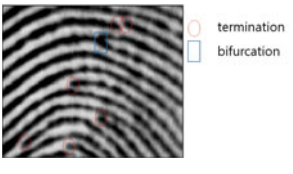


Fig. 2. Highlighted ridge termination and bifurcation minutiae

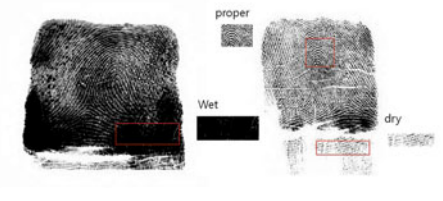


Fig. 3. Fingerprint humidity

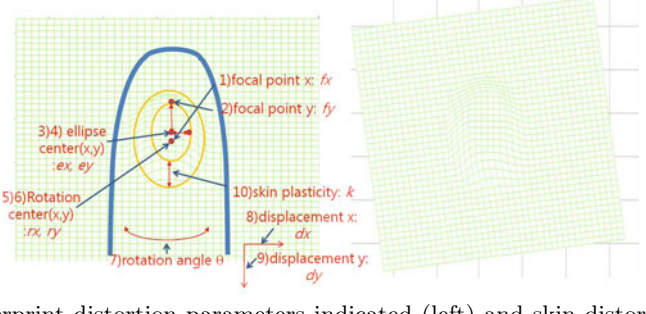


Fig. 4. Fingerprint distortion parameters indicated (left) and skin distortions [6] (rotation and translation) model example (right)

2.2 Distorted Image Collection

Four sources for fingerprint images include crime scenes, citizen databases, distortion collections, and training materials [4]. A fingerprint collection would be incomplete without numerous examples and combinations of classifications and minutiae. Our efforts presented here are to develop a DL for distorted images to be used in identification and matching algorithm analysis services for DLs that support the three other types of image sources.

A large set of distorted images from a known image is needed in order to test search algorithms. For the distortion parameters displayed in Fig. 4, the number of distorted images with full compositions of distortions will grow at a rate of 10^k, where k is the number of distinct levels selected for each distortion. For example, if $k = 5$, the levels of -8, -3, 2, 7, 12 pixels might be selected for x-axis displacement, and over 100,000 distorted images would be produced from the full-factorization of the ten parameters. Note that the number of levels could vary between distortion factors. Initial images in a distorted image collection come from human fingers, toes, palms, and foot pads. Assuming prints from entire palms and foot pads could be captured in one large image, a single human would produce 24 initial print images. Distorting each print over only five levels for the ten parameters leads to a sizable collection of 2,400,000 images per human.

Due to the extreme number of possible distorted images derived from a single human, especially as k grows large, high-quality algorithms need training to identify matches. A distortion collection is needed to build an algorithm testing platform with known fingerprint matches to test matching precision, recall, false negatives, false positives, and conditions that lead to low levels of certainty. Such a collection is useful in analysis to determine the contribution of each distortion in order to provide feedback to prediction algorithms. The collection we have developed uses a small set of initial images to allow for a larger number of distortion parameter levels. A system for managing distorted image collections also may determine which factors contribute to negatively impact matching algorithms. In order to conduct such analyses, each distorted image and sub- or partial image is stored with its generation parameters, provenance, and relationships to the unaltered initial image and other distorted images. Storage space plays a major factor in developing image collections. Currently, we have limited k to 5 as a terabyte of distorted images are generated per ten initial images.

3 Analysis and Experiment

To support a typical fingerprint algorithm analysis workflow, we have developed a DL services model and implemented a prototype instantiation. The workflow includes five stages: image harvesting, distortion image generation, algorithm execution, result harvesting, and algorithm performance analysis. With this model, researchers are able to investigate how an algorithm performs with synthetic, field-quality images. In particular, researchers are provided with an analysis framework that could be used to determine which image distortion parameters effect feature identification. The targeted workflow model framework pairs new

algorithms with image collections to allow analysis on which image characteristics effect algorithm performance, as seen in Fig. 5. Previous work has defined the format for formally defining DL services that we will employ [2]. Formal definitions already exist for generic versions of several of our fingerprint-specific! services, e.g., searching and visualizing [2]. The services in the following section are implemented in the prototype's initial workflow.

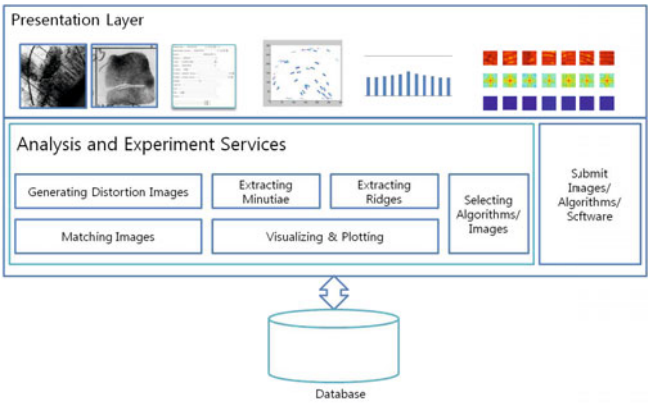


Fig. 5. Analysis and Experiment Services in DL framework

3.1 Analysis and Experiment Services

Analysis and Experiment Setting. Each algorithm to be used in experimentation requires an algorithm-specific description of which outputs to process for analysis. The minutiae extraction algorithm requires analysis on the number, pixel location, and quality of identified minutiae. The ridge tracing algorithm requires quantification and analysis on ridge identification. The matching algorithm requires analysis to determine if the set of selected minutiae was sufficient in making a match. One required step to begin an experiment is to set up all the input parameters using the 'Selecting algorithm/images' service to choose which images and algorithms should be selected in an experiment run.

Distortion Generation and Image Processing. After high-quality real-world fingerprint images are submitted, a distortion generation algorithm takes a set of values for the ten distortion parameters used in [6]. Each distorted image then is added to the DL collection along with a link to the original image and the parameters used in the image generation. In environments with low amounts of disk storage available, the image generator may be used to generate distorted images on a just-in-time basis, and images may be removed after usage though their metadata or generation script is archived.

For the image processing service, there are two approaches to the extraction of minutia and ridge features: *binarization* and *gray scale*. *Binarization* approaches typically use image processing techniques such as sharpening, histogram equalization, and enhancement, while *gray scale* approaches often exploit filtering

with a Gabor filter to enhance gray scale fingerprint images. This service can be considered as a converting service, similar to the distortion service, and thus has the same 5S formalization [2]. See Table 1 for basic terms and definitions.

Table 1. Basic Terms and Definitions of 5S formalization [2]

Term	Definition	Term	Definition
DO_i, DO_j	digital objects $i, j \in C$	V	Vertex
C	a collection $\in$ $Coll$	Stm_i	$\Psi_{ij}.Dom$
$Coll$	a set of collections	$\Psi_{ij}.Dom$	V $\times$ Streams
stm_j	a stream	S^3	Streams $\cup$ Structures $\cup$ Spaces
st_j	a structure	tfr	$S^3\times$ Spaces
Ψ	V $\times$ Streams $\Rightarrow$ (N $\times$ N)	sp_j	a space j
St^2	a set of functions Ψ		

Informally, *distorting* and *image processing* take a digital object and produce a distorted version by changing its streams, structures, or structured streams as defined in the 5S framework [2], an alternative to the DELOS reference model[1].

Definition. *Distorting/image processing* is a service defined as $f : do_i \Rightarrow do_j$, given a digital object do_i. The input and output structures for this service are do_i and do_j. The pre-condition and post-condition for this service are $\exists C \in Coll : do_i \in C$ and $\exists C \in Coll : do_j \in C$.

Minutiae Extraction and Ridge Tracing. A minutiae extraction algorithm is used to identify the locations and quality of major features, e.g., ridge bifurcation and termination. A third algorithm attempts to automatically trace the ridges in images resulting from smears, partial-smudges, or high humidity. High humidity refers to an overly oily or wet print that causes ridges to run together.

Definition. These two feature extraction algorithms form a service, *extracting*, that can be informally defined as *given a digital object, produce a descriptor from the object that represents the digital object.* User input is required as stm_i and outputs are (st_j, Ψ_{ij}). Pre-condition and post-condition are $stm_i \in Streams$ and $st_j \in Structs; \Psi_{ij} \in St^2; stm_i \in \Psi_{ij}.Dom; st_j.V \in \Psi_{ij}.Dom$, respectively.

Matching and Searching. A fourth algorithm for *matching* and *searching* attempts to use 3, 6, or 9-point triangles of high-quality minutiae locations to identify matches between two images as groups of minutiae are less susceptible to distortions. This matching algorithm stems from attempts to reduce the effects of small distortions on the identification of minutiae location and quality.

Definition. This process can be defined as a *binary* operation service $f(do_i, do_j) = k$, $k \in R$, compared to a service such as rating and measuring which is a *unary* operation $f(do_i) = k, k \in R$, where a real number k is a similarity score.

[1] See www.delos.info

Evaluating. Evaluation is a critical service among these experimental services. Evaluation criteria can be 1) algorithm performance, 2) algorithm efficiency, 3) minutia reliability, and 4) image quality. First, performance metrics include indicators used in the FVC such as 1) number of rejected fingerprints during enrollment; 2) number of rejected fingerprints during genuine matches; 3) number of rejected fingerprints during impostor matches; 4) impostor and genuine score distributions; 5) FMR(t) / FNMR(t) curves, where FMR is the false match rate, FNMR is the false non-match rate, and t is the acceptance threshold; 6) ROC(t) curve, where ROC is a receiver operating characteristic; 7) equal-error-rate (EER), the value that EER would take if the matching failures were excluded from the computation of FMR and FNMR (EER*); 8) the lowest FNMR for FMR$\leq$1%; 9) the lowest FNMR for FMR$\leq$0.1%; 10) the lowest FNMR for FMR=0%; and 11) the lowest FMR for FNMR=0%. Second, metrics for measuring efficiency include 1) average enrollment time, 2) average matching time, 3) average and maximum template size, and 4) maximum amount of memory allocated. Third, minutia reliability and image quality can be measured with the combination of many different maps (e.g., low frequency, high frequency, and directional).

Definition. Given a digital object, an evaluating service produces an evaluation (i.e., a real number) for it. Input is do_i and output is (do_i, w_i). Pre-condition is $\exists C \in Coll : do_i \in C$ and post-condition is $w_i \in [a, b] \subset R$.

Visualizing and Plotting. Analysis results can be visualized by projection to measurable spaces. Visualization techniques can be used to analyze the appearance and disappearance of minutiae over distortion degrees.

Definition. Visualizing and plotting can be described as processes that, given a collection, produce visualizations such as charts, histograms, plots, or meshes. Input for a visualizing service is a collection C and a transformation k, and output is a space j. Pre-conditions and post-conditions are $C \in Coll$ and $tfr_k(C) = sp_j \in Metric$.

3.2 Example Experiment Scenarios

Using these services, we have carried out three pilot experiments to analyze the effects of distortion on image quality. These experiments were performed regarding *fingerprint sufficiency* to provide objective standards of image quality.

Matching Score Accuracy Experiment. The first experiment analyzed the effects of skin distortion, specifically rotation and x-axis and y-axis displacements, on matching scores. This experiment followed the path: *distorting* $\Rightarrow$ *minutia extraction* $\Rightarrow$ *matching* $\Rightarrow$ *analysis* $\Rightarrow$ *visualization*. An example experiment for y-axis displacement is illustrated in Fig. 6. These led to another experiment to answer: "how does the distortion (e.g., rotation and translation in the skin distortion model) affect the number of minutiae?"

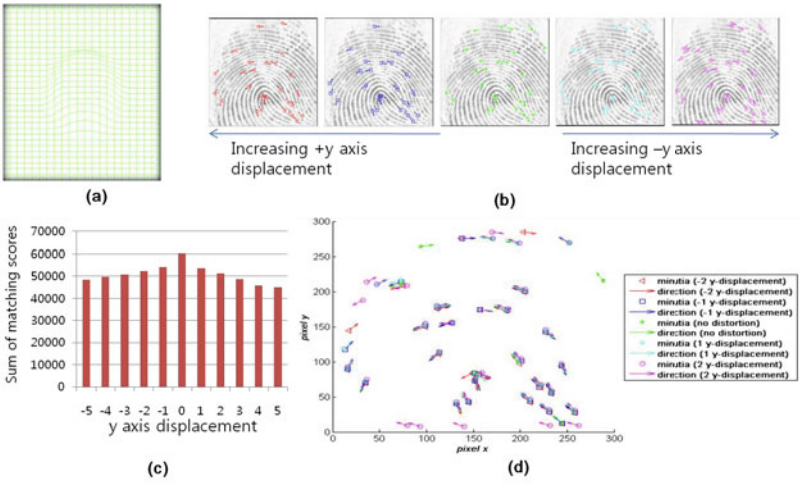


Fig. 6. Example analyzing effects of y-axis displacements on matching quality: (a) skin distortion model selected; (b) distorted images; (c) histogram of y displacement versus sum of matching score; (d) plotting of minutiae spatial distribution

Minutiae Count and Reliability. This experiment investigates the effects of distortions on minutiae count. We hypothesized that it would diminish the number of minutia points. For this experiment, we required minutia extraction services. This experiment was accomplished by following this path: *distorting* ⇒ *minutia extraction* ⇒ *analysis* ⇒ *visualization*. Fig. 7 shows that the number of minutiae increased as distortion increases.

Therefore, we investigated the effects of distortion on minutiae reliability, using the same experiment path used for the minutiae count test. In this case, minutia reliability was extracted and graphed to show the effects of distortion on extraction. In Fig. 8, the x-axis represents the amount of translation distortion ranging from -5 to +5 pixels for the x-axis and y-axis of a given image. (0,0) means no distortion in any image's axis, (0, -1) means y-axis distorted by -1. The y-axis represents reliability scaling from 0 (not reliable) to 100 (very reliable). We observed that average reliability and minutiae distortion are inversely related.

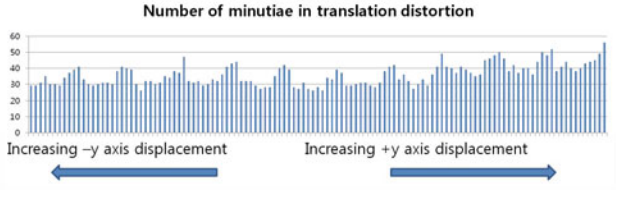


Fig. 7. Number of minutiae in translation distortion

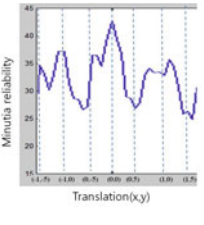


Fig. 8. Average minutiae reliability: images distorted by translation

Minutiae Plotting on Fingerprint. Plots for minutia reliability of each image follow the path: *distorting* ⇒ *minutia extraction* ⇒ *analysis* ⇒ *visualizing*. The results shown in Fig. 9 revealed that distortions have introduced false positive minutiae with low reliability, degrading the average minutia quality. The red points on the fingerprint image indicate low quality minutia points, whereas green points indicate high quality minutia points. The scale assigns red the lowest reliability and blue the highest.

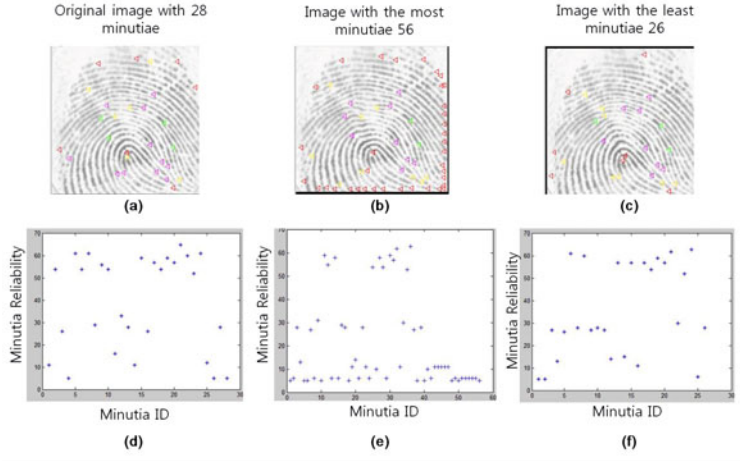


Fig. 9. Spatial distributions of minutia reliability: (a) original image with 28 minutiae; images which increased (b) and decreased (c) the most in minutia points after distortion (56 and 26 minutiae respectively); and (d)-(f) minutia reliability of each image

4 Framework and Prototype

We have designed a framework for workflowing image-based services, considering Kepler [5] and related existing systems.

4.1 Workflow, Experimentation, and Analysis Framework

When designing an image-based experiment, a user selects a collection of images and an algorithm to execute. Each algorithm in our framework is paired, by the algorithm developer, with an algorithm-specific analysis script to identify and extract the phenomenon being tested from the algorithm output. Our experimentation workflow involves executing each algorithm with a specific collection, e.g., a set of distorted images. After termination of the algorithm, the analysis service determines the differences in the algorithm results between the images. Metadata, such as distortion parameter values, then can be used to analyze each parameter's effect on the algorithm's results. Thus, the framework consists of building workflows of compositions of collections, algorithms, and analyses.

As an example workflow, the number of minutiae located and the assigned quality score (0.0 to 1.0) for each minutiae is provided by the minutiae extraction algorithm. An analysis script extracts the Cartesian location and assigned quality scores. The results from executing this algorithm on the entire set of distorted images from a base image then is matched with distortion parameters. The predefined analysis script is executed to specifically determine the statistical significance of each factor in hindering the identification of minutiae. This algorithm requires the distortion generation algorithm as a pre-requisite, forming a workflow involving several algorithmic executions and subsequent analysis.

4.2 Prototype Implementation

We have implemented a basic prototype of this framework to conduct experiments with the feature extraction algorithms previously mentioned. The prototype consists of DL services to manage a distorted image collection, select and execute an algorithm, and execute analyses. The analysis processes allows a researcher to hold several parameters constant by careful selection of distortion parameters, e.g., x-axis translation, rotations, and skin plasticity. We are developing a plug-in system for easier integration of new algorithms.

We have developed a collection of real-world images. For several selected images, we have generated a range of distorted images and produced a service for generating new distorted images as required. An experiment was successfully designed, executed, and analyzed to determine the effects of humidity, x-translations, y-translations, rotations, and skin plasticity on minutia extraction.

The prototype includes an online collection of original and distorted images and a system for selecting and composing service workflows. The Google chart API is used to present results of completed analysis tasks. A web-interface is used to browse the image collection, image information, distortion parameters used to generate specific images, extracted minutiae, and ridge information.

Currently, our prototype system contains 137,785 prints (FVC2000: 3520, FVC2002: 3520, SD27: 516, self-collected: 629, and distorted: 129,600). For the preliminary experiments, we generated distorted images from real fingerprints as described in Section 2.2. As a result of our experiments, the system yielded the following: 1) matching scores of a minutia extraction module MINDTCT and BOZORTH3 produced by National Biometric Image Sofware of BIST matching algorithm with distorted image sets (see Figure 6); 2) minutia counts of MINDTCT algorithm with distorted image sets (Figure 7); 3) minutia reliability of MINDTCT algorithm with distorted image sets (see Figures 8 and 9); and 4) improvement of schema presented in previous work [4].

Our framework is scalable but limited by file system storage space. Current terabyte storage devices can have roughly 1-10 billion images assuming 800 base images and 100,000 distortions at 100KB per file. Distorted image generation time is 1.0 sec. on a Pentium 4. Both time and space complexity are $O(n)$.

4.3 Other Applications

This framework, developed for fingerprint testing and analysis, can be applied to a broader set of domains. Hays and Efros [3] recognize the location pictured by a given image by extracting visual features of that image and matching them to a previous known geo-tagged image database. The most similar geo-tagged images retrieved tell at which point on Earth the given input image is located. Retrieval algorithms for this type of task are quite precise. A researcher might test different algorithms to check their sensitivity to distorted or rotated images using the proposed framework.

The problem of astronomical identification resembles fingerprint matching in the sense that once stars are extracted from an input image, they are matched to a guide stars catalogue. A great number of algorithms for this problem have been proposed, but as in the fingerprint domain, the precision of star identification is influenced by extraction of a star centroid and preprocessing of the raw image [7]. The study of how those factors influence the precision of proposed algorithms, similar to triangle minutiae matching, can benefit from this framework as well.

5 Related Work

FBI's Integrated Automated Fingerprint Identification System (IAFIS) is a large fingerprint management system, supporting search capabilities against both latent and ten prints, storing electronic images, and electronically exchanging fingerprints. However, it does not support a series of services for experiment digital libraries such as experiment setting, distorting, plotting, and visualizing. The Universal Latent Workstation (ULW) is the first latent workstation supporting interoperability and sharing latent identification services with local and state authorities, and with the FBI IAFIS, all with a single encoding.

Penatti et al. [9] proposed an experiment management tool, *Eva*, for evaluating descriptors in content-base image retrieval, providing image descriptors, and image management, to run comparative experiments. This tool has stimulated the development of our holistic DL experiment framework. Previous work also supported scientific communities in a web-based integration framework [10].

Fingerprint analysis has been challenged by various distortions such as merged prints, pressured impressions, humidity on fingertips, partial prints, or simultaneous prints. Distortions are likely to affect minutia extraction quality, ridge tracing quality, matching scores, and image quality. The Analysis, Comparison, Evaluation and Verification (ACE-V) and Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST) groups (see swgfast.org) have worked on fingerprint analysis. Oliveira et al. [8] proposed a multiscale directional operator and morphological tools for reconnecting broken ridges in fingerprint images. Huang et al. [1] proposed singular point detection.

From the object perspective in very large digital libraries, Kozievitch et al. [4] proposed an solution to integrate four different very-large fingerprint digital libraries. A proposed compound object (CO) scheme uses the 5S framework, modeling different types of objects found in those DLs, to allow uniform use

in an integrated DL. Our work is focused on designing a DL framework, from a services perspective, to deliver analytical results of an experiment that integrates related services designed by different researchers.

6 Summary and Future Work

Our main contribution is supporting collaborative research for researchers and trainers with services for generating distorted image datasets, testing different algorithms (e.g., for minutia detection and matching), and managing and workflowing scientific research datasets, algorithms, and analysis results.

We are integrating each implemented service under the proposed framework. We plan to verify this prototype in terms of algorithm correctness before and after integration. In addition, we will confirm that findings of experiments relate to the practice of researchers and fingerprint analysts. We also plan to incorporate (training and matching) algorithms from three other types of fingerprint DLs [4] with our collection of distorted images. Astronomy and geo-location identification domains provide a parallel corpus of algorithms that compare images based on feature extraction. Comparisons of these algorithms would be useful for cross-domain generalization.

Acknowledgments. We are grateful for NIJ (Award No. 2009-Dn-BX-K229), CAPES (BEX 1385/10-0), and BAE Systems support.

References

1. Huang, C.-Y., Liu, L.-m., Hung, D.C.D.: Fingerprint analysis and singular point detection. Pattern Recognition Letters 28(15), 1937–1945 (2007)
2. Gonçalves, M.A.: Streams, structures, spaces, scenarios, societies (5s): A formal digital librarY Framework and its applications. Ph.D. thesis, Virginia Tech, Blacksburg, VA, USA (2004)
3. Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2008)
4. Kozievitch, N., da Silva Torres, R., Fox, E., Park, S.H., Short, N., Abbott, L., Misra, S., Hsiao, M.: Rethinking fingerprint evidence through integration of very large digital libraries. In: Ioannidis, Y., Manghi, P., Pagano, P. (eds.) Proceedings of the Third Workshop on Very Large Digital Libraries VLDL 2010. Institute of Information Science and Technology of the National Research Council (ISTI-CNR), Pisa, Italy (2010)
5. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience 18(10), 1039–1065 (2006)
6. Maltoni, D., Cappelli, R.: Advances in fingerprint modeling. Image and Vision Computing 27(3), 258–268 (2009); special Issue on Multimodal Biometrics - Multimodal Biometrics Special Issue

7. Na, M., Jia, P.: A survey of all-sky autonomous star identification algorithms. In: 1st International Symposium on Systems and Control in Aerospace and Astronautics (ISSCAA 2006), pp. 896–901 (January 2006)
8. Oliveira, M.A., Leite, N.J.: A multiscale directional operator and morphological tools for reconnecting broken ridges in fingerprint images. Pattern Recognition 41(1), 367–377 (2008)
9. Penatti, O.A., da Siva Torres, R.: Eva: an evaluation tool for comparing descriptors in content-based image retrieval tasks. In: Proceedings of the International Conference on Multimedia Information Retrieval, MIR 2010, pp. 413–416. ACM, New York (2010), http://doi.acm.org/10.1145/1743384.1743455
10. Wang, F., Rabsch, C., Kling, P., Liu, P., Pearson, J.: Web-based collaborative information integration for scientific research. In: IEEE 23rd International Conference on Data Engineering, pp. 1232–1241. IEEE, Los Alamitos (2007)

A Novel Combined Term Suggestion Service for Domain-Specific Digital Libraries

Daniel Hienert, Philipp Schaer, Johann Schaible, and Philipp Mayr

GESIS – Leibniz Institute for the Social Sciences,
Lennéstr. 30, 53113 Bonn, Germany
{Daniel.Hienert,Philipp.Schaer,Johann.Schaible,
Philipp.Mayr}@gesis.org

Abstract. Interactive query expansion can assist users during their query formulation process. We conducted a user study with over 4,000 unique visitors and four different design approaches for a search term suggestion service. As a basis for our evaluation we have implemented services which use three different vocabularies: (1) user search terms, (2) terms from a terminology service and (3) thesaurus terms. Additionally, we have created a new combined service which utilizes thesaurus term and terms from a domain-specific search term recommender. Our results show that the thesaurus-based method clearly is used more often compared to the other single-method implementations. We interpret this as a strong indicator that term suggestion mechanisms should be domain-specific to be close to the user terminology. Our novel combined approach which interconnects a thesaurus service with additional statistical relations outperformed all other implementations. All our observations show that domain-specific vocabulary can support the user in finding alternative concepts and formulating queries.

Keywords: Evaluation, Term Suggestion, Query Suggestion, Thesaurus, Digital Libraries, Interactive Query Expansion.

1 Introduction

A general and long known problem with keyword-based search is the so called "vocabulary problem" or "wording problem" [6]. The same information need or search query can be expressed in a variety of ways. Current web search engines often retrieve a list of documents where same relevant items are always included – but this is mostly a phenomenon of the very large document index. Thus, when using a "wrong" term there is still a high probability getting a non-empty result set.

When we analyze today's Digital Library (DL) systems or domain-specific databases a controlled vocabulary, usually a thesaurus is used to index the publications. DLs often consist of metadata entries on the specific publications, descriptive abstracts are optional. In this situation the vocabulary problem can become quite serious. If the searcher doesn't use one of the controlled terms the document was indexed, the chance of getting relevant documents is low. There is a significantly higher chance to retrieve an empty result set. Users tend to adapt their search strategies to work around these drawbacks. In a user

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 192–203, 2011.

study done by Aula et al. [1] one expert articulated: "I choose search terms based not specifically on the information I want, but rather on how I could imagine someone wording […] that information."

Modern information-seeking support systems (ISSS) try to make use of a variety of automated approaches to transform and expand textual queries e.g. by using stop word lists, stemming or spelling correction. From the perspective of interface design interactive query reformulation still is an open research issue. [11].

In the following paper we will present the results of a user study with more than 4,000 unique visitors in the online information portal Sowiport[1]. Users were confronted with three basic term suggestions services based on (1) user-search-terms, (2) terms from a terminology service and (3) terms from a domain-specific thesaurus. As a novel approach, we have created a term suggestion service that combines thesaurus terms and terms from a domain-specific search term recommender. We will present related work in section 2, followed by the evaluated vocabularies and services in section 3. We will proceed with the conducted evaluation in section 4 and will present results in section 5. We conclude this paper with a discussion in section 6.

2 Related Work

We will present two different perspectives on query reformulation tools: the origin of the proposed terms and the different types of reformulation tools.

As Efthimiadis [5] points out interactive query expansion (IQE) can be divided in two types of IQE mechanisms: (1) those that are based on collection dependent or independent knowledge structures and (2) those that are based on the search results. The difference between these approaches is the origin of the data to propose terms from. The terms that are presented to the user can either be retrieved from a knowledge structure like e.g. thesauri or from the documents that are included in the search result (e.g. to perform a pseudo-relevance feedback). Regarding this characteristics Vechtomova et al. [18] compared two approaches for query expansion (QE) based on term co-occurrences. The first approach was a global co-location analysis where the entire document collection was used to extract related terms. The second approach only used terms from a local subset of the retrieved documents. This local approach clearly performed better than the global one. The difficulties in the first global approach seemed to lie in proposing too unspecific and too general terms [3]. The authors argued that users need to have a more context specific QE mechanism.

The fact that users need supporting mechanisms to correctly formulate their queries is supported by the user studies of Hargittai [7]. She found that 63% of the participants made a typographical or spelling mistake of some kind, and among these, 35% made only one mistake, but 17% made four or more errors during their entire session. This is supported by the search engine logs analysis of Cucerzan and Brill [4]. They found that 10 – 15% of the queries had typographical error.

Hearst [8] described two types of supporting: spelling suggestions/corrections and automated term suggestions. Term suggestion can be further differentiated in pure term and query suggestion like shown by Kelly et al. [10]. While a term suggestion is

[1] http://www.gesis.org/sowiport

only focused on single terms, a query suggestion tries to combine suggested terms with other terms and present them to the user as a new and complete query. Query suggestions therefore can provide alternative viewpoint and can help to explore unfamiliar scientific areas. Like Kelly showed users preferred the query suggestion method and rated it higher. This included the ability to help them think about new approaches for their search.

Query and term suggestions are implemented in many modern web search engines but many systems only try to make suggestions based on prefix matches (user types "soc" and the system suggests "social", "society" and so on) while the actual origin of the suggestions remains unclear. A typical representation of this kind of suggestion was shown by White and Marchionini [20]. They performed a study on an interactive method, which they called "real time query expansion". After the user types a word and presses the space bar, the system presents terms based on the surrogates of the ten top-ranked documents. On these prototypes White et al. [19] conducted a usability study with 36 participants, each doing two known-item tasks and two exploratory tasks, and each using the baseline system, the query suggestions, and two other experimental interfaces. For the known-item tasks, the query suggestions scored better than the baseline on all measures ("easy", "restful", "interesting", etc.). Participants of their study were also faster using the query suggestions over the baseline on known item tasks and made use of the query suggestions 35.7% of the time. Those who preferred the query suggestion interface said it was "useful for saving typing effort" and "for coming up with new suggestions".

A more general study on Web search interfaces was performed by Jansen et al. [9]. They studied a search engine log file from Dogpile.com with 2.5M interactions (1.5M of which were queries) from 2005. Using their computed session boundaries (mean length of 2.31 queries per session), they found that more than 46% of users modified their queries, 37% of all queries were parts of reformulations, and 29.4% of sessions contained three or more queries.

Regarding the combination of these approaches Schatz et al. [16] did a study on two different term suggestion methods: one using terms from a subject thesauri and the other from term co-occurrence lists. The overall finding was that multiple approaches resulted in a better search quality. They suggest combining different IQE methods and origins in favor of a single method.

In the following sections we will describe our different term suggestion services and the set-up of our evaluation.

3 Term Suggestion Services

We implemented three basic term suggestion services with different vocabularies as a basis for our evaluation: user-search-terms (UST), a terminology service (HTS) and a social science thesaurus (TS). Additionally, as a novel approach, we have created a combined term suggestion service (CTS) which combines the social science thesaurus and a search term recommender service.

3.1 Basic Term Suggestion Services

We use different vocabularies as a data basis for the basic term suggestion services in our study that are introduced in the next sections: (1) User-Search-Terms (UST), (2) Terminology Service (HTS), and (3) Thesaurus Terms (TS).

The service UST is an uncontrolled set of terms extracted from the query log of the social science portal Sowiport. We recorded about 28,000 distinct terms entered by human users since 2007. The applied service includes all user terms from the query log as a flat list of terms without any additional information (see Fig. 1a). The terms are chosen by matching the input term against a ranked list of the user terms. The user terms are ordered by the frequency count of their usage.

The service HTS is a controlled set of terms coming from a terminology service (called heterogeneity service) implemented in the portal Sowiport [12]. The service contains controlled terms from 25 different thesauri with about 26,500 distinct terms. The thesauri are connected with intellectually created relations that determine equivalence, hierarchy (i.e. broader or narrower terms), and association mappings between terms. To search and retrieve terminology data from the heterogeneity service an individual has to enter correct terms from at least one controlled vocabulary of the service. For the recommendation service we used an adapted heterogeneity service that returns a flat list of related terms ordered alphabetically (see Fig. 1b).

The Thesaurus for the Social Sciences (Thesaurus Sozialwissenschaften) is an instrument to index and retrieve subject-specific information in Sowiport. The list of keywords contains about 11,600 entries, of which more than 7,750 are descriptors and about 3,850 are non-descriptors. Topics in all of the social science disciplines are included. The applied service includes all descriptors from the thesaurus as a flat list of terms without any relational data and ordered alphabetically (see Fig. 1c).

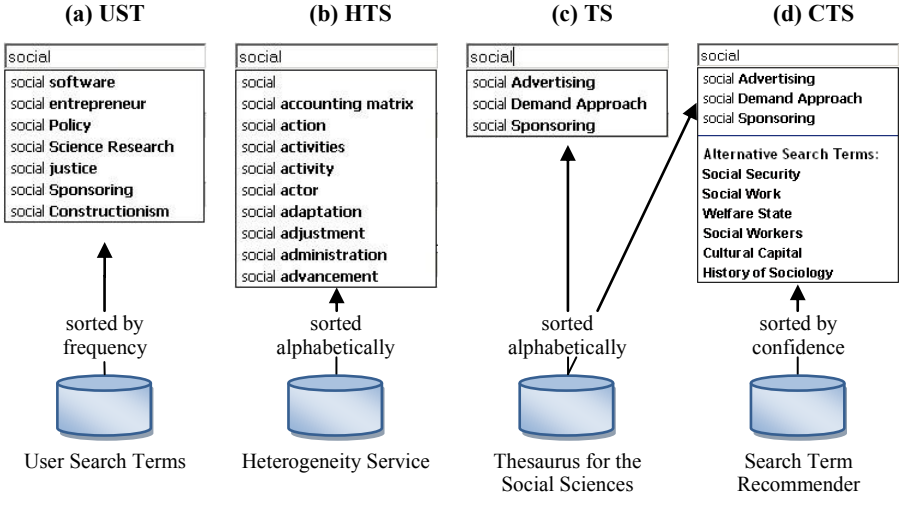


Fig. 1. Four implemented term suggestion services: (a) User-Search-Terms (UST), (b) intellectually mapped terms (HTS), (c) Social-Science Thesaurus (TS) and (d) Combined Term Suggestion (CTS)

3.2 Combined Term Suggestion Service

As a data basis for our Combined Term Suggestion Service we use thesaurus terms and terms from a Search Term Recommender (STR). Petras [14] proposed a search

term suggestion system, which relies on two basic parameters: (1) the controlled vocabulary terms that are used for document representation and (2) the natural language keywords that are input by the searcher. The advantage of suggesting controlled vocabulary terms as search terms is that these terms have been systematically assigned to the documents, so that there is a high probability of relevant and precise retrieval results if these terms are used instead of whatever natural language keywords the searcher happens to think of.

The STR addresses the problem of search term vagueness by performing a co-word analysis of the terms of a field in order to recommend more appropriate terms to the user. The STR maps query terms to indexing terms at search time by building term-term-associations between two vocabularies: natural language terms from titles and abstracts on the one hand side and controlled vocabulary on the other hand side. The associations are weighted according to their co-occurrence within the collection to predict which of the controlled vocabulary terms best mirror the search terms.

In the original implementation Plaunt and Norgard [15] used a likelihood ratio statistic to measure the association between the natural language terms from the collection and the controlled vocabulary terms to predict which of the controlled vocabulary terms best mirror the topic represented by the searcher's search terms. Given a training set of documents containing free terms form title/abstract and controlled vocabulary terms a dictionary of co-words can be build which includes strength of association (the calculated weight). This can be used to predict the possibility of likelihood between words.

Our own implementations rely on latent semantic analysis and support vector machines. They are applied via the commercial indexing software Mindserver. The used service returns a flat list of terms corresponding to the input term ordered by the strength of association.

The combined term suggestion service combines the TS service with recommendations from the search term recommendation service. Until three characters it shows terms from the thesaurus, beginning with four characters it shows an additional section with *Alternative Search Terms* under the TS list. We used the limit of four letters to avoid input terms for the STR that leads to poor results. Term suggestions that appear in both lists are filtered out and are shown only in the TS section.

4 Evaluation

In this section we first present the social sciences information portal Sowiport as a real-world environment for our user study. In section 4.2 we describe the logging process and the evaluation periods.

4.1 Evaluation Environment

We chose the social science information portal Sowiport as a real-world environment for our user study. Sowiport integrates literature references, persons, institutions, projects, services and studies. It currently contains about 4.8 million literature references and research projects from 18 databases, including six databases from ProQuest/CSA, which are available by a national license funded by the German Research

Foundation. The German-language share of the databases include the GESIS own databases SOLIS and SOFIS, which contain about 500,000 literature references and research projects which are indexed intellectually with the Thesaurus for the Social Sciences. Sowiport is offered in German and English, the majority of users are from German-speaking countries. The portal reaches about 7,000 unique visitors per month. The term suggestion functionality has been integrated in the simple search form on the home page and in the advanced search form. The term suggestions are proposed to the user as a list under the input field. The user can choose a term from the list with a mouse click or by scrolling and return. The term is then entered into the input field. With a click on the button *Search* or with *Enter* the search is submitted.

4.2 Logging Process and Evaluation Periods

For conducting the user study, we had to log the entered search terms, the selection from the recommendation's list and the search queries performed. Taking this information from the server log can be a very time consuming and error-prone issue. Server logs are full of records from irrelevant search engines and crawlers and we want to make sure only to log data from human users. We therefore implemented a function that logs all this information only if a user clicks on the search button or hits enter in the search field. In particular we have logged the following information: (1) for a selection of a recommendation from the list: entered term, chosen term, position of the chosen term, service type, date/time and session id. (2) For a submitted search: submitted term, date/time and session id.

The different vocabularies UTS, HTS, TS and CTS were sequentially activated in Sowiport in a time period of about 3 months. Each service was activated until the count of visitors using the search had exactly reached 1000 unique visitors. Once the number was reached we changed the service to the next one. A unique visitor is identified technically by an internal ID. The user can perform several actions like browsing or searching in the database, but is recognized only once as a unique user. A user session is still valid for two hours if the user performs no further actions.

5 Results

In this section we will show the individual results of the conducted user study. In section 5.1 we will present results of the use of term suggestion services and in section 5.2 we will show a categorization of patterns we have found in the data.

5.1 Use of Term Suggestions Services

We used the unit measure of 1000 unique visitors to calculate the share of selected recommendations to all users. This number describes the average use of search term suggestions based on unique users. For the CTS approach 50.9% of the users used the recommendation service, followed by the TS vocabulary with 37.5%, the UST vocabulary with 25.2% and the HTS with 10.4%.

As a second measure we have calculated the share of selected recommendations to all searches performed. This number describes the use of recommendations based on

all searches and shows therefore the general use of search term suggestions. The CTS performed best with 14% usage, then the TS vocabulary with 9%, the UST with about 7% and the HTS service with only about 3% usage. The number has been in all cases under 15%, which means, at best, only in one of seven search queries the recommendation service has been used, in the case of the HTS service only in three of one hundred searches. In general we can say that there is a very weak use of the recommendation service, one would expect a much higher usage. Possible reasons, such as the count of letters entered to the choice of term, the word length of the chosen term etc. are discussed in the following paragraphs.

On the basis of the collected data we were able to calculate the average position of the selected term from the recommendations list and the average count of letters entered to the choice of the term. On average the selected term had the second position on the list. Figure 2 shows the distribution of the position of the chosen term in the list. The graphs of the individual vocabularies and services show a similar trend. The percentage of selection of the concept at first to tenth position decreases, this means, recommendations on a high position in the list are chosen clearly more often.

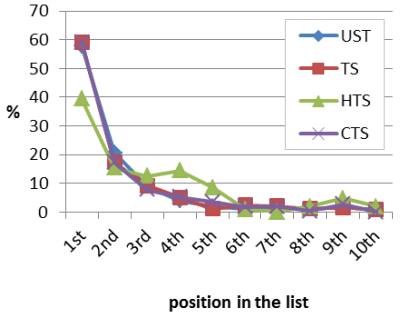


Fig. 2. Rank of the chosen term in the list of recommendations

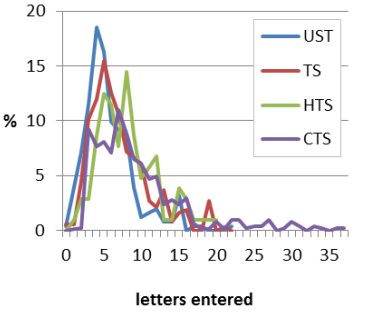


Fig. 3. Number of letters entered until a term is chosen

The user enters about nine letters, before he chooses a term. The average word length of the chosen concept has been 15 letters, including terms and term combinations. This is more than twice the average word length of German words of 6.44 [13]. This means the user selects very long terms and term combinations. Even with excluding terms combinations, the average word length of individual chosen terms with about 14 letters is still very high. The distribution of the word length of the entered term in figure 3 shows peaks from 6 to 11 letters. To summarize, this leads to the following conclusions:

- Recommendations were used at best for every seventh search query.
- Recommendations on a high position in the list are chosen clearly more often.
- For terms with short and normal word length recommendations are not selected.
- In contrast, terms and term combinations with very long word length are selected.

Table 1. Summarized results: key figures for each technique, their use in relation to unique users and search queries and measures for entered and chosen terms

	UST	HTS	TS	CTS
Unique users	1000	1000	1000	1000
Search queries	3566	3572	4165	3604
Selected recommendations	252	104	375	509
Share of selected recommendations to all searches	7.06%	2.91%	9%	14.12%
Share of selected recommendations to all unique users	25.2%	10.4%	37.5%	50.9%
Average position of the selected concept	2	2.9	2.1	2.1
Average count of letters entered to the choice of the term	6.8	9.2	7.7	11
Average word length of chosen concepts (single terms and combinations)	15.1	16.6	14.8	15.2
Average word length of single terms	13.2	15.6	13.6	13.6

5.2 Patterns of Use

The data indicates different patterns of use, which can be classified in different categories. We analyzed the evaluation data, identified different categories for the transition from entered to chosen term and classified them into these categories. We found four different categories for all services, two more categories for the User-Search-Terms vocabulary and one category for statistically near terms of the CTS approach. The different categories are:

1. *Simple term completion*: the user enters initial letters and chooses a concept from the list.
2. *Selecting an already completely entered term*: the user enters a concept completely and then selects the same concept from the list.
3. *Selecting an already completely entered term, after a simple term completion in the search before*: in the first search the user enters initial letters and chooses a concept from the list, in the following search the user enters the complete concept and then chooses the same concept from the list.
4. *Term extension*: the user enters a complete term and then chooses a concept with term extension from the list.
5. *Two complete terms entered, second one changed*: the users enters a concept with two terms completely and then chooses a concept from the list where the second term changes.
6. *Selecting a more abstract concept*: the user enters a fine-grained concept and then chooses a more abstract concept from the list.
7. *Statistically near term*: the user enters a term and chooses a statistically near term form the *Alternative Search Terms*-section of the CTS approach. In contrast to the other categories, the terms have no common stem between entered and chosen term, but are statistically near.

Table 2. Examples for the different categories

Category	Entered term	Chosen term
1	acci	accident
2	accident	accident
3	acci accident	accident accident
4	accident	accident analysis
5	cognitive maps	cognitive development
6	mother-child clinic	mother
7	medicine	Doctor-patient-relationship

Table 3. Frequency of categories in different services with more than 2%. For CTS individual results for each section and total results for the whole service are shown.

Category	UST	HTS	TS	CTS	
1	52.98%	52.89%	64.27%	49.71%	2.35%
				52.06%	
2	9.92%	16.34%	13.6%	13.75%	0.4%
				14.15%	
4	36.11%	30.77%	20.53%	4.9%	6.29%
				11.19%	
7					22%
				22%	

Simple term completion is used in more than 50% of cases (TS: 64.27%, HTS: 52.89%, CTS: 52.06%, UST: 51.98%) in all different services. The high ratio in the TS vocabulary indicates that here the proposed terms most likely correspond to the ones the user thought of. Selecting an already completely entered concept from the list is a pattern that occurs regularly: 16.64% in the HTS, 14.15% in the CTS, 13.6% in the TS and 9.92% for the UST vocabulary. The explicit selection from the list might be a cognitive assurance coupled with an action that chooses a concept in the sense of a controlled vocabulary. The user might think that if the system proposes a concept, it must be correct even if one enters the term himself before. The selection of an already entered term, after a simple term completion in the search before occurred six times only in the TS vocabulary. Here the user might have learned that (1) the concept exists and (2) that it leads to (useful) results. (1) is proven to the user through the appearance of the concept in the recommendations list, (2) is proven to the user by already seeing the result list for this search term. Term extension is the second big ratio for term completion: 36.11% for the UST vocabulary, 30.77% for the HTS, 20.53% for the TS and 11.20% for the CTS.

Categories 5 and 6 are patterns that occurred only within the User-Search-Term vocabulary. In category five the user enters a concept consisting of two terms and then chooses a concept from the list where the second term is changed, what happens three times. Selecting a more abstract concept has been a very rare occasion with only

two examples: entered term: “mother-child clinic”/selected term: “mother” and entered term: “antidiscrimination eu”/selected term: “antidiscrimination”.

Category 7 exist only for the CTS service and contains statistically near chosen terms, which are not simple term completions or extensions with common stem to the entered term, but are statically near terms based on co-word analysis. The number of 22% for the whole service therefore represents the number that could only be achieved by this particular service.

6 Discussion

In this study we evaluated an interactive term suggestion service for the domain-specific DL Sowiport. The service was tested with three different vocabularies (UST, HTS and TS) and a combination of TS and STR.

Term acceptance was generally comparable to other studies dealing with term suggestion methods, where the thesaurus-based method clearly scored best. Here the acceptance rate was between 37.5% and 50% (see table 1) which can be compared to other user studies in this field [19, 20]. The thesaurus-based method clearly outperformed the other single-method implementations which is a strong indicator that term suggestion mechanism need to be domain-specific to gain acceptance from the users.

This can be explained with the “Anomalous State of Knowledge” [2] wherein the user is while formulating queries. In this state he tries to map the words and concepts describing his problem to the terms of the system while typically fighting ambiguity and vagueness of language. This problem especially occurs in highly specialized scientific literature databases where often literature reference with spare bibliographic metadata is available for matching.

In scientific communities special discourse dialects evolve. These dialects are not necessarily the same dialects an information specialist or user would use to describe a document or a concept using a documentation language. The consequence is a serious source of vagueness in the query formulation phase. A term suggestion method per-se can support the user in this early stage of the search process e.g. choosing an appropriate query term which is used in the language of documentation.

It can be easily seen that different implementations of term suggestion services match different suggestion tasks: The simple term-completion tasks (category 1) are best matched by the TS, while near terms (category 7) are only suggested by the STR. UST can best match the need to extend terms with different concepts (category 4). While the categories 1 and 7 can be explained with their immanent features to be an expression of the language of indexation (in case of the TS) and an expression of the language of discourse (in case of the STR), the UST represents the language of the user looking for information. Here we can see the unfiltered search terms and queries users are actually using.

User studies in digital libraries have shown that most users are not aware of the special controlled vocabularies used in digital libraries [17]. Hence they are not using them in their query formulation. This can be seen in our relatively low acceptance rates of 9% of the thesaurus-based implementation. To overcome the acceptance problems we derived the plan to combine the rather simple term but most accepted completion method (TS) with a more sophisticated term-mapping (STR) approach.

We can observe a significantly increase of the acceptance rate because these approaches complete each other.

On the one hand the TS service contains terms which are very relevant to the domain but on the other hand it is quite limited in its ability to deliver suggestions for all possible user aspects. This is due to its size of around 7,750 terms which have to map after three entered characters. In contrast the STR can nearly always suggest a controlled term since it maps a total of 1.45 million free terms on the 7,750 TS terms. The overall advantages of suggesting controlled vocabulary terms as search terms is that these terms have been systematically assigned to the documents, so that there is a high probability of relevant and precise retrieval results.

Our paper, especially the CTS approach, introduces new possibilities of suggestions in domain-specific DL. In the case of the STR we can see that term suggestions could provide a broader overview over different areas of a scientific domain or discussion, which typically involves particular associated concepts (perhaps assuming different meanings or directions of thought). The result is a diverse domain perspective on certain concepts, an effect that can also be achieved by displaying the semantic term mappings themselves. The general assumption of our paper is that term suggestions from several fields of research can provide a new view or different domain perspective on a topic in an interactive way. Combining different specific term suggestion methods is not an academic exercise; quite on the contrary, our approach has been clearly confirmed by users in a large scale real-life scenario.

References

1. Aula, A., Jhaveri, N., Käki, M.: Information search and re-access strategies of experienced web users. In: Proceedings of the 14th International Conference on World Wide Web, pp. 583–592. ACM, New York (2005)
2. Belkin, N.J.: Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science 5, 133–143 (1980)
3. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Inf. Process. Manage. 43, 866–886 (2007)
4. Cucerzan, S., Brill, E.: Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP (2004)
5. Efthimiadis, E.N.: Query expansion. Annual Review of Information Systems and Technology (ARIST) 31, 121–187 (1996)
6. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The Vocabulary Problem in Human-System Communication. Commun. ACM 30(11), 964–971 (1987)
7. Hargittai, E.: Hurdles to information seeking: Spelling and typographical mistakes during users' online behavior. Journal of the Association of Information Systems 7(1), 52–67 (2006)
8. Hearst, M.: Search User Interaces. Cambridge University Press, Cambridge (2009)
9. Jansen, B.J., Spink, A., Koshman, S.: Web searcher interaction with the Dogpile.com metasearch engine. J. Am. Soc. Inf. Sci. Technol. 58, 744–755 (2007)
10. Kelly, D., Gyllstrom, K., Bailey, E.W.: A comparison of query and term suggestion features for interactive searching. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 371–378 (2009)

11. Marchionini, G., White, R., Belkin, N., Golovchinsky, G., Kelly, D., Pirolli, P., Schraefel, M.: Information Seeking Support Systems: An invitational workshop sponsored by the National Science Foundation (2008)
12. Mayr, P., Petras, V.: Cross-concordances: terminology mapping and its effectiveness for information retrieval. In: 74th IFLA World Library and Information Congress, Québec, Canada (2008)
13. Nettle, D.: Segmental inventory size, word length, and communicative efficiency. Linguistics 2, 359–367 (1995)
14. Petras, V.: How one word can make all the difference - using subject metadata for automatic query expansion and reformulation. 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. 21–23. Springer, Heidelberg (2006)
15. Plaunt, C., Norgard, B.A.: An association-based method for automatic indexing with a controlled vocabulary. J. Am. Soc. Inf. Sci. 49, 888–902 (1998)
16. Schatz, B.R., Johnson, E.H., Cochrane, P.A., Chen, H.: Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval. In: Proceedings of the First ACM International Conference on Digital Libraries, pp. 126–133. ACM, New York (1996)
17. Shiri, A., Revie, C.: Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. J. Am. Soc. Inf. Sci. Technol. 57, 462–478 (2006)
18. Vechtomova, O., Robertson, S., Jones, S.: Query Expansion with Long-Span Collocates. Inf. Retr. 6, 251–273 (2003)
19. White, R.W., Bilenko, M., Cucerzan, S.: Studying the use of popular destinations to enhance web search interaction. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 159–166. ACM, New York (2007)
20. White, R.W., Marchionini, G.: Examining the effectiveness of real-time query expansion. Inf. Process. Manage. 43, 685–704 (2007)

Did They Notice? – A Case-Study on the Community Contribution to Data Quality in DBLP

Florian Reitz[1] and Oliver Hoffmann[1,2]

[1] University of Trier, Universitätsring 1, Trier, Germany
reitzf@uni-trier.de
[2] Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Germany
hoffmann@dagstuhl.de

Abstract. Defective metadata is a significant problem of digital libraries. So far, automatic error detectors have been in the focus of research interest. However, recent public projects have shown that patrons are willing to invest time to report errors if they are called to contribute. In this case-study, we analyze the community contribution to error detection for DBLP, a public bibliographic collection. Our study is based on e-mails sent to the project between January 2007 and November 2010. We manually and automatically identify error reports and analyze their contribution to corrections of the DBLP collection. We show that users frequently report certain types of defects while others are ignored. The detection of homonym-name inconsistencies in particular strongly depends on user input. We also discuss who sends the reports and which communities are particularly active in this matter.

1 Introduction

One of the most important tasks that digital library projects face every day is assuring the quality of metadata records for the entities they store. Threads to the quality range from missing, incomplete or outdated information to data which is misleading or simply wrong [10]. Defective metadata have numerous negative effects on the performance of a digital library. Consider the DBLP bibliographic collection which stores information on more than 1.6 million publications. The major quality issue of DBLP is the identification of researchers which – lacking a better system – has to be done by personal names. The personal name is a poor identifier because it is neither unique (several persons have the same name) nor stable (persons change their names). A wrong identification can cause publication lists to get split or mixed up with the publications of other persons.

There are many algorithmic approaches which have been proposed to detect and remove inconsistent data. The person-name problem in particular has drawn some attention. Solutions range from learning algorithms designed to detect inconsistencies (e.g. [4][9]) to clustering approaches which generate person entities by applying similarity or dissimilarity metrics to the citations (e.g. [3][5]). A point which has so far been neglected are the contributions users make to data quality. There are several open digital libraries which allow their patrons to edit metadata records or which provide interfaces which allow error reports from inside the application [2]. For example, the United States Library of Congress published a set of images on Flickr with the request

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 204–215, 2011.

for corrections and additional information. Responses to this call were considered to help improve the data quality [14]. DBLP, on the other hand, is a closed system with an interface which does not provide interaction. However, many patrons have the urge to support the project by pointing out defects. The only way to report a defect is by sending an e-mail. In this study, we analyze e-mails which were sent to DBLP between January 2007 and November 2010. We show that a significant part of these mails actually reported inconsistencies and that many reports caused modifications to the collection. The main contribution of the paper is a case study of error reports and their significance to a widely used digital library. To this goal, we also consider the persons who reported the defects.

This work is structured as follows: In Section 2 we discuss the underlying data sets which we use for our study including an evaluation of mails sent to DBLP. In Section 3 we describe the significance of the reports and analyze the properties of the reported defects. In Section 4 we present results on the persons who submitted the reports and on the importance of communities to the amount of contribution. We conclude our work with a discussion of related work.

2 The Data Sets

This work is based on two data sets. One contains all modifications to DBLP since 1999 and the other is a corpus of mails received by the DBLP project between January 2007 and November 2010. At first, we discuss corrections to DLBP in general. Then we examine the mail corpus with special attention to the way defects were reported.

2.1 Changes and Corrections in DBLP

DBLP grows fast, but besides the newly added publication records there is also a number of changes to the existing ones. Each publication record is an XML element which consists of sub-elements or fields that hold the actual metadata. `<author> John Doe </author>`, for example, stores one author name. A *change* to a record can either add, remove or change a single field. We denote the value of the field before and after the change as *old* and *new* respectively. Some modifications require several changes. For example, when author *J. Doe*, who has 10 papers, is renamed to *John Doe* there are 10 similar changes – one for each publication record. If these changes are performed on the same day we call them a *correction*. We obtain past changes and corrections from the hDBLP data set[1]. From this file, we can reconstruct the DBLP data set for each day since June 1999. For further details on the hDLBP and the extraction of changes and corrections, please refer to [12]. Between January 2007 and December 2010, we recorded 186,464 changes of which 60,238 (32%) added, 28,126 (15%) removed and 98,100 (53%) modified fields. We also considered changes and corrections from December 2010 to capture delayed fixes of error reports in the last mails from November of the same year. The changes group to 122,983 corrections. Tables 1(a) and 1(b) show the number of changes and corrections we recorded for the different types of fields. For comparison, *occurence* lists the average number of these elements in DBLP during the considered period.

[1] http://dblp.uni-trier.de/xml/hdblp.xml.gz

Table 1. Overview on changes and corrections between January 2007 and December 2010

(a) changes

field	occurence	changes	wait
author	3,150,610 (34.3%)	99,865 (53.6%)	45,073
ee	795,374 (8.7%)	38,872 (20.8%)	29,242
note	13,027 (0.1%)	12,570 (6.7%)	1,480
title	1,432,869 (15.6%)	10,138 (5.4%)	201,813
url	1,148,850 (12.5%)	7,550 (4.0%)	216,764
others	2,372,346 (28.8%)	17,469 (9.4%)	239,995

(b) corrections

field	occurence	avg. changes
author	70,546 (57.4%)	1.42
ee	38,132 (31.0%)	1.02
note	1,149 (0.9%)	10.94
title	2,226 (1.8%)	4.55
url	7,544 (6.1%)	1.00
others	3,386 (2.7%)	5.41

Not all types of fields are changed with the same frequency. More than half of all changes alter an *author* field which underlines the significance of the personal name problem. *ee* is a reference to an electronic edition of a paper and usually points to a publisher's web site. *url* stores a link to a researchers web site. These links are listed on the top of some author pages. In general, changes are rare events. Column *wait* lists the expected number of days we have to wait before a specific data field in a specific publication record is changed – assuming that corrections are evenly distributed. We will use this fact later to identify changes reported by e-mails. For corrections, the fraction of *author* is even higher. *avg. changes* lists the average number of changes which make up a correction.

2.2 E-Mails

The mail corpus contains messages to the DBLP project received between January 2007 and November 2010. Most spam and personal messages were removed but the corpus still contains mails which are no error reports. The majority of mails are written in English or German. We excluded 23 mails which we could not open. We ignored any message part which is not plain text. This includes attached PDF files but not HTML mails. To make sure that we did not miss important information, we manually checked 100 random attachments but found no error reports. Overall, we retained 6311 mails, 1580 written in 2007, 1705 in 2008, 1664 in 2009 and 1362 in 2010.

To learn more about the mail corpus, we conducted a manual examination of 1000 randomly selected messages. In this sub-set, we found 458 messages which were intended as error reports. At this stage, it is not relevant whether a mail actually caused modifications to DBLP. We also consider mails which request adding or updating a link as an error report. Name-related inconsistencies are the most important type of reported errors. There are 217 mails reporting a *synonym* – one person is listed with different names so there are multiple publication lists for one researcher. A *homonym* – a name which refers to more than one person so publication lists contain the work of different writers – is reported 92 times. 28 mails report more than one and up to nine single defects.

DBLP does not specify a format in which error reports may be submitted. Not surprisingly, there is a wide variety of styles in the mail corpus. Below, we see four typical examples generated from real e-mails. They all differ in the quantity of given personal name information (underlined) and citation metadata (printed in bold).

m_1	Please have a look at authors John Doe and John A. Doe they seem to be the same person and should be merged. Thanks J.
m_2	There is a paper **My title** published in **2005** at **venue** by **J. Doe**. The year is wrong, it should be 2006.
m_3	Please have a look at paper **conf/venue/XYZ09**. I think the author names are wrong.
m_4	There is an error in my 2nd publication. Please have a look at it, John

We considered the way in which persons and citations are referred to. This is important as the user is not guided by policies or assisted by user interface. There are different ways to refer to a person. Each person in DBLP has a unique key which can be found in the URL of the respective publication list. For *Florian Reitz* the key is *Reitz:Florian*. The name itself can be given as full (*Florian Reitz*) or in abbreviated form (*F. Reitz*). In the order *key – full – abbreviated*, the names become more ambiguous. If a report contains person entities 50% of them are identified by the person key and 74% by the full name. Note that DBLP stores abbreviated names for many publications so using them is not necessarily a result of a careless submitter.

References to citations are less frequent than references to persons and can be found in 44% of the reports. This is not surprising given the large number of name-related defects which are not always related to a single publication. As for persons, DBLP defines keys for citations. If a report contains a citation the most popular information is the combination of *title + authors* (61.7%). Citation keys are less frequently used than person keys. They can be found in 30% of reports with citations. Person keys appear in the URL of author pages while viewing citation keys requires to follow a link. This might make them less perceivable for the user. Kapoor et al. [6] noted a similar reluctance to use unique identifiers – DOIs and ISBNs in their case – when storing citations. However, in the mail corpus, title and authors are sufficient to uniquely identify all citations. Most submitters preferred to cite publications by using a text fragment containing all relevant metadata. Machine readable metadata were seldom used, mostly in form of HTML fragments and BibTeX. Most HTML fragments were copied from the DBLP web page. No report used the DBLP internal XML representation for citations.

A more significant problem is the use of relative information. In mail m_4, for example, it is not clear what exactly the *2nd publication* is. We found references like this in 60 mails. Most common are the DBLP publication numbers which we detected in 40 mails. These numbers enumerate the publications on an author page. These numbers are not stable. As a result of this evaluation, we are currently planning to implement a more restrictive error submission web site. On the one hand, it will force the submitters to give more detailed information but on the other hand, it will also assist them by proposing complete metadata or personal names.

To get information on the whole mail corpus and prepare further steps of our analysis, we automatically examined persons and citations in all mails. We used hDBLP to generate two authority lists which contain the names of all persons listed in DBLP and all citations respectively. With these lists we searched for person and citation entities in the corpus. Table 2 lists the number of detected entities for the specified settings. We

Table 2. Identified person and citation entities with different filters

(a) person

	category	detected	mails
P_1	*full*	14252	4618 (73.2%)
P_2	*partial*	8389	2853 (45.2%)
P_3	*abbreviated*	37051	3430 (54.3%)
P_4	*key*	3101	1783 (28.3%)
P	union	59656	5619 (89.0%)

(b) citation

	category	detected	mails
C_1	*publication key*	1215	423 (6.7%)
C_2	*title + full name*	5305	1155 (18.3%)
C_3	*title + partial name*	138	106 (1.7%)
C_4	*title + abbrev. name*	98	45 (0.7%)
C_5	*title + name key*	1464	561 (8.9%)
C_6	*title + year + pages*	4453	980 (15.5%)
C	union	6973	1636 (25.9%)

ignored names which are part of other names like *Michael Le* $\subset$ *Michael Ley* as well as very frequent names. In addition to abbreviated names, we also consider *partial* names which are like full names without middle name parts and name extensions like *sen.*. The first column of each table gives the name of the setting. The results confirm the findings from the manual examination. Abbreviated names are highly ambiguous. We will not use them in our study because they produce a high number of erroneously identified person entities.

3 The Reports

To get a global view on the user contributions, we must determine whether a mail is an error report or not. The manual evaluation covers only a fraction of the mail corpus. To get information for all messages, we apply an automatic identification algorithm based on the person and citation entities we automatically detected. We then analyze the reports and discuss their significance to the quality management of DBLP and the properties of the repaired defects.

3.1 Automatic Detection of Error Reports

To find reports, we apply a simple heuristic. Consider mail m_1 from Section 2.2 which reports a name-related inconsistency. Fixing this inconsistency requires changes to author fields which are related to person entities *John Doe* or *John A. Doe*. If we can find such changes in the set of all modifications we call them *triggered by* m_1. We examine modifications which can be detected up to 30 days after the mail was received. Consider Table 1: the expected waiting time for all types of fields is well above 30 days, i.e., with high confidence all changes we find are related to this mail. We call a correction *triggered* by a mail if at least one of its changes is triggered. For detected citations, we examine all changes to fields of the respective metadata record. For a person p, we consider all records of publications authored by p but we register only changes which have the name of p as old or new value. There are two reasons for doing so: (1) Some persons have many publications so we examine changes from a large number of fields which increases the chance of random hits. (2) If only a personal name is given, it is likely that the defect is related to the person and not to a specific publication. Fixes

of name-related problems result in changes as described above. We also examine any modification to the author's personal record.

We detected 7996 triggered changes which is 4.3% of all changes in the period January 2007 to December 2010. 7486 changes were detected by considering person entities and 1536 by considering citation entities. The total number for persons is not much higher than the number for citations because most citations contain personal names as well. The number of identified persons is much higher than the number of citations so the larger number is not surprising. The number of triggered corrections is 6195 which is 4.91% of all corrections. The average waiting time between receiving the report and the change is 1.28 days if a change is triggered by a citation and 1.52 days if it is triggered by a person. In any case, most reports were processed during the first 24 hours after being received.

Of all mails, 47.2% triggered changes and corrections which is slightly more than we obtained from the manual analysis. 77 mails which were listed as reports (16.9%) did not trigger changes. Of those mails 44 actually contained information on defects which should have triggered changes. We do not know why these mails were not considered. 17 mails were meant to be reports but the requested corrections were defective. 16 mails provided insufficient information for the automatic detector. 68 mails labeled as non-report caused modifications (12.5%). 80% of the non-reports triggered changes to the author record, i.e., they added or modified homepages or affiliation information. These data came from the mail signatures and were obviously updated as a side effect. Five mails were replies to earlier reports and 8 mails contained large lists of citation data. It is likely that a change can be found for a large list.

3.2 Are Reports Significant for DBLP?

The fraction of triggered changes and corrections to the total number of changes and corrections seems to be small on first sight. If we consider them more closely we find that the distribution of field type differs from the distribution we see in Table 1(a). For example, 53.6% of all changes affect the *author* field. For the triggered changes, the share of this field type is 66%. Figure 1 shows the combination of field types for the different sets of detected entities. *All* is the union of sets P and C. The *author* field

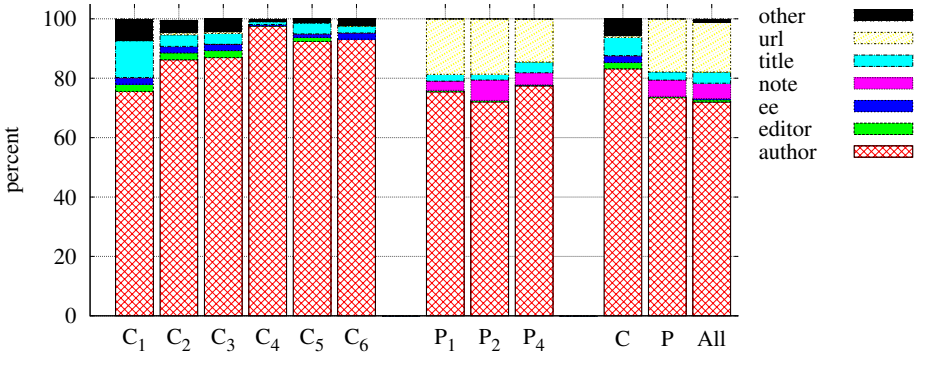


Fig. 1. Affected field types for different filters (for filter names see Table 2)

is particularly frequent for the citation sets. For C_4 – the most ambiguous category of citation entities – 97% of all changes affect the *author* field. These citations are listed with abbreviated names. For abbreviated names, name-related inconsistencies are much more likely than for citations with full names. As a result, name defects are more frequent and their corrections are usually more urgent so there is a high number of reports on this field. We assume that the high number of triggered *author* changes is a result of the high visibility of related defects. If an *author* field is defective the respective paper will appear in the wrong publication list. This is far more eye-catching than other defects like a wrong page number. Changes to *ee* – the second most frequent field type – are almost never triggered by mails. We considered changes to this field more closely: most changes were done en masse. More precisely, 95% of all changes to the *ee* field were done on just 65 days. The *ee* field is poorly suited to be reported by mail. Usually, if a change to a citation of a venue is required all citations must be altered as the web location for all of them has changed. This type of modification is best done in batch by using some update script.

The uneven distribution of error reports affects the individual recall values for different types. Table 3 lists the fraction of reported corrections to the total number of corrections for the most common field types. For most fields, the results deviate from the average value. The *url* field in particular has a high contribution. *ee* on the other hand is almost never reported. The table also shows the fractions for each year in the study. We registered the maximum for 2008 where 8.3% were triggered followed by the minimum in 2009 where this was only the case for 3.65%. However, the number of triggered corrections is about stable.

Table 3. Fraction of triggered corrections by type for different years

	author	**editor**	**ee**	**note**	**title**	**url**	**others**	
2007	776 5.11%	9 8.11%	6 0.16%	32 24.43%	20 13.25%	270 13.16%	16 6.15%	1129 5.24%
2008	1225 7.97%	20 19.61%	17 0.65%	148 9.55%	27 3.18%	424 27.37%	27 3.66%	1888 8.30%
2009	1098 5.59%	14 22.58%	6 0.03%	127 5.50%	62 8.42%	401 17.36%	26 3.02%	1734 3.65%
2010	969 4.76%	8 4.68%	5 0.05%	103 26.28%	489 21.47%	1633 14.88%	11 1.02%	1444 4.20%
	4068 5.77%	51 11.43%	34 0.09%	410 9.36%	214 9.61%	1338 17.74%	80 2.72%	6195 4.91%

We see that the majority of induced corrections is related to the *author* field. If we consider these corrections with respect to the homonym/synonym problem more closely we find four relevant types [12]: **merge**: removes a synonym by unifying all related *author* fields to a single name. **split**: removes a homonym h by changing some author fields with content h so that the publications stored for h are now divided upon two persons. **distribute**: reassigns some author fields of a person p to another person p' which already existed. Before this correction, p was homonym and synonym at the same time. **rename**: Consistent renaming of a person p. p was neither homonym nor synonym. For

the period between January 2007 and December 2010, we detected 28,151 mergers, 2,658 splits, 10,920 distributions and 13,228 renamings. Among the triggered corrections are 443 distributions, 1,693 mergers, 226 renamings and 417 splits. The mergers, distributions and renamings make up 6.0%, 4.1% and 1.7% of all corrections respectively. Triggered splits cover 15.7% of their category. Usually, a synonym or homonym problem is related to a name but not to a single publication. Therefore, it is not surprising that corrections related to name-inconsistencies are far more often induced by personal names than by citations. The low number of renamings is surprising as they can not be discovered automatically. We considered renames more closely and found many cases where abbreviated names were changed to full names. This increase of information is usually provided by publishers and not by users.

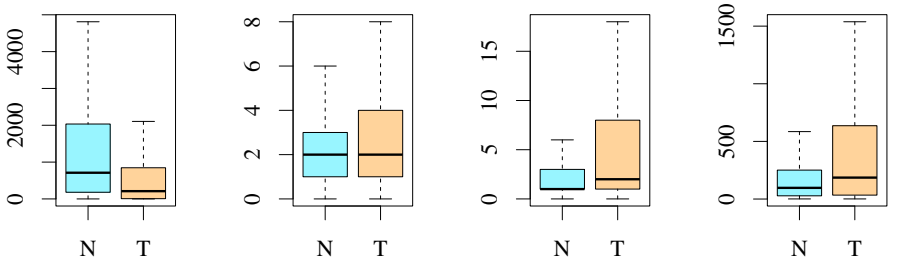
Another class of defect is missing or surplus data. This type of defect is difficult to detect automatically because similarity or dissimilarity metrics do not work here. We detected 3253 triggered changes which added a field and 588 which deleted one. The fraction of all changes is 5.4% and 2.09% respectively. 17.87% of added *author* fields were triggered as were 3.1% of the removed.

3.3 Do Triggered Changes Differ from Others?

Now we can compare properties of triggered and non-triggered changes. At first, we examine the affected citation records. In general, triggered changes affect younger citations than others. Figure 2(a) shows a box plot for the absolute age of citations, i.e., the number of days between adding the record to DBLP and the modification. If the change is triggered the mean citation age is only half as high as if the change is not triggered. Only few non-triggered changes affect citations which are younger than 100 days. The median for triggered changes however is only slightly higher (211). Obviously, new citation records are more interesting for the users. We assume that many persons wait for their citations to appear and check them immediately. In Section 4.1 we will see further evidence for this. We expected to find more triggered changes for citations with many authors. In this case, a defect appears on many author pages. Though there is a tendency in this direction, this effect is small (Figure 2(b)).

Next, we consider name-related changes. Figure 2(c) shows differences in the number of publications[2]. The median and mean values of publications are twice and 3.8 times as high respectively for triggered changes as median and mean for non-triggered. Persons with many publications are usually more central to the field and therefore defects are detected faster. We examine 200 randomly selected persons with a single publication who were affected by non-triggered changes. We find a high number of scientists who published several years ago. We will see that many reports are sent by the person affected by a defect. Few people with no recent contribution to computer science would check DBLP for defects. To get a real estimation on the popularity of a person we examined the number of times the respective author page was viewed by users. Figure 2(d) shows the number of page impressions logged on the main DBLP server between October 2007 and October 2010. Again, persons affected by triggered changes are more popular than others. However, 6.9% of these persons have ten or less page impressions.

[2] $min = 0$ because of persons which were added during a change and had no publication before.



(a) age (days) (b) authors (c) publications (d) page impression

Fig. 2. Box plots of different properties for entities affected by triggered changes (T) and non-triggered changes (N). Outliers are not displayed.

4 The Submitters

In a final step, we examine the persons who submitted the 2976 automatically identified error reports. Identifying these persons is not easy. E-mail addresses are usually unique but a person can have more than one. We tried to parse signatures which can be found at the end of many mails but there are many different designs and we were not able to extract reliable data. For this study, we use the sender name – an optional field which is provided with the mail address. We found 2268 distinct submitters. 1654 of these submitters are listed in DBLP as authors or editors. Note that we could not identify all submitters so the number of those who are listed in DBLP is probably higher. However, there is also a noticeable number of non-scientific staff or young PhD students who hand in reports.

4.1 Who Submitted Reports?

There are two extremes of user contribution frequency: (1) all reports are sent by a small group of heavy submitters, (2) many submitters sent a single message each. DBLP tends towards the second type. Table 4 lists how the submitters are distributed by number of reports. 83% of the submitters handed in a single report. There are only 11 heavy submitters who sent more than five reports. Note that the most frequent submitter sent 137 reports (5%). All heavy submitters are listed in DBLP except for one. This person is a non-scientific worker who was charged with keeping the records of a work group clean. Figure 3 shows how the reports are distributed among the groups of submitters. Together the heavy submitters sent 8% of the relevant messages. Having a large number of infrequent submitters is positive as the project does not want to depend on a single person. On the other hand, we noted that the quality of reports is higher for frequent submitters. All heavy submitters usually used citation or person keys if possible and – with the exception of one person – avoided the problems we discussed in Section 2.2.

Table 4. Frequency of submitters

reports sent	persons	listed DBLP
1	1890	1357 (71.8%)
2	275	218 (79.3%)
3-5	92	69 (75.0%)
6-20	10	9 (90.0%)
137	1	1 (100.0%)

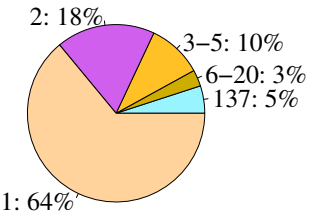


Fig. 3. Fraction of reports

We found that many error reporters are interested in the correction of defects which are related to themselves. This is not surprising: DBLP has become an important instrument to assess the work of scientists so there is a pressure to assure completeness and correctness of one's own entry. Authors also have a more detailed knowledge of their work which makes it easier to spot defects. 1364 reports (45.8%) triggered a change to the name of the submitter, 1502 reports (50.5%) altered a publication authored by the submitter, 480 times, we observed a change to the name records of a coauthor of the submitter. Altogether 1836 (61.7%) reports triggered changes directly or almost directly related to the submitter. For the heavy submitters, the number of self-related messages is much smaller, namely 15.6%. This is not surprising as it is unlikely that there are that many defects related to a single person.

4.2 Interest of Communities

Computer science consists of several sub-fields. In prior work [11], we showed that for some of these fields a number of conferences is missing while others are almost completely represented in DBLP. We can assume that DBLP is less relevant for communities which are poorly covered than for those with high coverage. This might influence the number of triggered corrections we find for a sub-field. We apply a sub-field framework proposed by Laender et al. [7] in 2007 and refined by Martins et al. [8] in 2009. The framework features 27 sub-fields and a list of 1000 conferences. Each conference is assigned to one sub-field. Journals are missing from this list. We computed the percentage of triggered corrections to citations of a sub-field and compared the result with the total number of changes.

The percentage of triggered changes differs by sub-field. If we consider all changes it ranges from 1.2% (*Robotics Control and Automation*) to 7.8% (*Computational Biology*). The Pearson correlation between coverage of a sub-field and percentage of triggered changes is 0.252. If we only consider changes to the *author* field values range between 1.5% (like before) and 13.9% (*Databases, Information Theory ...*) for the original core-community of DBLP. The correlation between coverage and fraction of triggered name changes is 0.414. For non-name related changes the correlation is weaker (0.12).

5 Related Work

The idea of users contributing to metadata quality in general has been the subject of several studies. Since 2007 the United States Library of Congress has been publishing

pictures from their archive on the community platform Flickr. As for DBLP, users are not able to directly modify the metadata but they can post comments. The comments are similar to mails as they allow free text. Unlike the situation we analyze, the mapping of comment and picture is always clear and users can view past comments and react to them. Zarro et al. [14] analyzed the public response to this experiment. They evaluated the comments to 1043 pictures in three categories. For one category they found "Corrections and Translations" for 13% of pictures while for the others the fraction was at about 1%. The authors state that reports on wrong data "have the potential to be a great benefit of a commenting system". Other groups concentrate on technical issues of submitting reports. Bovey [2] presented an integrated interface which allows registered users of the Kent Cartoon Center project to modify document metadata. As for DBLP, all modifications need approval from an editor before they become visible to other users. The study does not contain results on user contributions.

To the best or our knowledge there has been no study on mail contributions to closed digital libraries like DBLP. However, there is similar research in the field of open-source project analysis [1][13]. Open source projects resemble digital libraries. They store source code documents which are organized by a hierarchical structure. Though the projects are usually open for direct contribution many users prefer to send corrections or extensions by mail as a *patch*. Weißgerber et al. [13] analyzed the mail contributions of two open source projects.

6 Conclusions

We presented a case study on e-mail-based error reports for the DBLP project. Our results can not easily be transferred to other applications. However, closed projects like DBLP – which do not actively invite user contribution or provide a respective interface – are common and user contribution might be relevant for them as well. In general, corrections caused by error reports make up only 4.9% of all corrections which is a small percentage. However, for some type of information the contribution is more significant. The name-inconsistency problem in particular benefits from the work of the users. For example, reports triggered 15.7% of all split modifications. Name-inconsistencies are particularly difficult to deal with and any improvement in data quality might lead to further corrections as the automatic approaches can work on a better data basis.

There are many persons who submit defects. Most of them are interested in correcting defects in their own work but almost 40% of the reports deal with unrelated defects. There are also contributions from persons who are not listed in DBLP though this is difficult to estimate. For DBLP we found that user triggered corrections deal with new defects while other corrections mostly clean older data.

References

1. Bird, C., Gourley, A., Devanbu, P.T.: Detecting Patch Submission and Acceptance in OSS Projects. In: Workshop on Mining Software Repositories, p. 26. IEEE CS, Los Alamitos (2007)

2. Bovey, J.: Adding User-Editing to a Catalogue of Cartoon Drawings. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 457–460. Springer, Heidelberg (2006)
3. Ferreira, A.A., Veloso, A., Gonçalves, M.A., Laender, A.H.F.: Effective self-training author name disambiguation in scholarly digital libraries. In: Hunter, J., Lagoze, C., Giles, C.L., Li, Y.-F. (eds.) JCDL, pp. 39–48. ACM, New York (2010)
4. Han, H., Giles, C.L., Zha, H., Li, C., Tsioutsiouliklis, K.: Two supervised learning approaches for name disambiguation in author citations. In: Chen, H., Wactlar, H.D., Chen, C.c., Lim, E.-P., Christel, M.G. (eds.) JCDL, pp. 296–305. ACM, New York (2004)
5. Han, H., Zha, H., Giles, C.L.: Name disambiguation in author citations using a K-way spectral clustering method. In: Marlino, M., Sumner, T., Shipman III, F.M. (eds.) JCDL, pp. 334–343. ACM, New York (2005)
6. Kapoor, N., Butler, J.T., McNee, S.M., Fouty, G.C., Stemper, J.A., Konstan, J.A.: A Study of Citations in Users' Online Personal Collections. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 404–415. Springer, Heidelberg (2007)
7. 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)
8. 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: Heath, F., Rice-Lively, M.L., Furuta, R. (eds.) JCDL, pp. 193–202. ACM, New York (2009)
9. On, B.-W., Lee, D., Kang, J., Mitra, P.: Comparative study of name disambiguation problem using a scalable blocking-based framework. In: Marlino, M., Sumner, T., Shipman III, F.M. (eds.) JCDL, pp. 344–353. ACM, New York (2005)
10. Redman, T.C.: Data Quality for the Information Age, 1st edn. Artech House, Inc., Norwood (1996)
11. Reitz, F., Hoffmann, O.: An Analysis of the Evolving Coverage of Computer Science Subfields in the DBLP Digital Library. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 216–227. Springer, Heidelberg (2010)
12. Reitz, F., Hoffmann, O.: Learning from the Past: An Analysis of Person Name Corrections in DBLP Collection and Social Network Properties of Affected Entities. In: Memon, N., Alhajj, R. (eds.) International Conference on Advances in Social Networks Analysis and Mining, pp. 9–16. IEEE Computer Society, Los Alamitos (2010)
13. Weißgerber, P., Neu, D., Diehl, S.: Small patches get in! In: Hassan, A.E., Lanza, M., Godfrey, M.W. (eds.) Workshop on Mining Software Repositories, pp. 67–76. ACM, New York (2008)
14. Zarro, M.A., Allen, R.B.: User-Contributed Descriptive Metadata for Libraries and Cultural Institutions. In: Lalmas, M., Jose, J.M., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 46–54. Springer, Heidelberg (2010)

A Comparative Study of Academic Digital Copyright in the United States and Europe

Robert J. Congleton and Sharon Q. Yang

Moore Library, Rider University, 2083 Lawrenceville Road, Lawrenceville, New Jersey 08648, U.S.A.
{rcongleton,yangs}@rider.edu
http://www.rider.edu

Abstract. The advent of Internet and digital media has added more complications to the already complex copyright laws. This paper will first summarize the history of copyright laws in the United States and Europe. It will then analyze and compare the digital copyright laws as they are applied in higher education in the United States and major countries in Europe.

Keywords: copyright, digital copyright law, United States copyright law, European copyright law, fair use; fair dealing.

1 Introduction

Copyright is defined as the legal protection given to the creator of an intellectual property. Such creators "are entitled to exclusive rights, such as the right to display, perform, transmit, and copy their own work and prepare new works based on the original work" [1]. Copyright law always tries to keep a balance between the interests and creativity of copyright holders and the advances of science and technology for the benefits of mankind. The ever changing nature of new digital media and use of the Internet to broadcast and access material threatens to tip the scale of copyright law in favor of either harsher restrictions or more open access. To address these balances copyright law must be constantly revised.

Modern copyright laws in the United States and Europe are based on several international agreements. The three primary international treaties are the Berne Convention for the Protection of Literary and Artistic Works of 1886, last revised in 1979, and two World Intellectual Property Organization treaties finalized in 1996: the WIPO Copyright Treaty (WIPOCT), and the WIPO Performances and Phonograms Treaty (WIPOPPT). Both WIPO treaties became effective in 2002. The Berne Convention recognized two important aspects of copyright: the moral rights of the author, and the need to have certain free uses of works for criticism, teaching and news. However it was left up to the individual signing nations to legislate how such exceptions will be permitted for educational use, "provided such utilization is compatible with fair practice" [2]. The WIPO Copyright Treaty clarified that reproduction rights and exceptions regarding computer programs, no matter in what mode or form they are expressed, are covered within the articles of the Berne Convention, while the WIPO Performances and Phonograms Treaty did the same for

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 216–226, 2011.

performances and recordings [3, 4]. The wording of each treaty allows for coverage to be extended to digitized or web-based modes or forms even though they are not explicitly mentioned. The treaties generated much controversy during their initial negotiations, and in the aftermath of the final agreements. These arguments carried over to heated public discussions as each signing government drafted legislation to meet the terms of the treaty. Some of the most important issues concerned the moral rights of the author, exceptions to copyright, anti-circumvention of technological protection measures (TPM), digital rights management (DRM), and suggested remedies for infringement.

This paper is a comparative study of copyright law and how it has been adapted by the United States, the European Union (EU), and selected European nations to address the needs of higher education in the digital age. The focus is on the use of digital materials in colleges and universities and how the laws govern educational uses in the United States as compared to Europe. Pertinent international treaties, relevant lawsuits and court decisions are discussed and philosophical differences between the approaches taken by the United States and Europe are described.

2 Digital Copyright for Higher Education in the United States

The first United State copyright law was written in the 1700s and covered printed media. It was revised many times over the years to address new media and drastically rewritten in 1976 under Title 17 of the United States Code. Under US copyright law, the creator of an intellectual property has exclusive rights to his or her work. The creator receives the copyright automatically without registering with the Copyright Office once his creativity appears in a tangible form or "fixed in a copy" [1]. The duration of the copyright comprises the creator's life time plus 70 years after his or her death. The basic principles of copyright law governing the printed materials remain true today with digital media.

The U.S. copyright law can be very prohibiting without fair use and other exemptions. Literally speaking, it is a violation of copyright if someone is "whistling a tune while walking down the street (public performance)" or "quoting from a novel in a review (reproduction)" [5]. Therefore to encourage the advances of science and technologies as well as teaching and research in other areas, the U.S. Library of Congress and the U.S. Copyright Office have made exemptions to allow the use of copyrighted materials under special circumstances without asking for the permission of the creator or paying royalties. One such exemption is called "fair use".

Section 107 of Title 17 in the United States Code is on fair use. The umbrella of fair use covers "criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research" [6]. Use and reproduction of copyrighted materials for the above purposes does not need permission from the owners and is not an infringement of copyright. Fair use is "by far the most enigmatic doctrine in U.S. copyright law and by far the most important. Without it, much of our economic and communicative action would constitute copyright infringement" [7].

It was not until the 1998 Digital Millennium Copyright Act (DMCA) that copyright law began to directly address digital issues. The new digital law incorporated the two 1996 treaties of the World Intellectual Property Organization

(WIPO): the WIPO Copyright Treaty and the WIPO Performances and Phonograms Treaty. It prohibits circumvention of anti-piracy measures (digital rights management systems) for accessing and copying digital works. However, circumvention for copying is allowed under fair use. An academic institution is not responsible for its faculty or student in copyright infringement unless it involves required materials for online courses or the institution has received two notifications for copyright infringement in the past three years. Nor are colleges and universities liable for infringement criminal charges. Under the new digital copyright law "the institutions must provide all of its users with informational materials describing and promoting compliance with copyright law" [8]. The US Copyright Office reviews the DMCA and revises exemptions every three years.

The new digital law was met with strong criticism and accused of favoring copyright holders. The critics argued that the new law seemed to forbid all circumvention of digital rights management systems including fair or non-infringing use of copyrighted works. In 2002 the U.S. Congress addressed these concerns with the passage of the Digital Choice and Freedom Act, and the Technology, Education, and Copyright Harmonization (TEACH) Act. The TEACH act has significantly affected education by redefining "the terms and conditions on when accredited, nonprofit educational institutions throughout the U.S. may use copyright protected materials in distance education-including on websites and by other digital means--without permission from the copyright owner and without paying royalties" [9].

The TEACH Act defines what distance learning is and gives detailed provisions about fair use in the online environment versus face to face teaching. For instance, the digital materials used in distance learning shall be limited to access by the class for the duration of the course. The institution must deploy technical measures that reasonably prevent the students from retaining the materials beyond the class as well as distributing them. The institution must form policies on and provide notice about copyright.

Fair use doctrine has not changed in the digital age. Academic institutions rely on fair use to teach, learn, and conduct research using copyrighted materials. However, the U.S. Copyright law does not offer a definition of fair use, but enumerates four broadly worded factors that courts and everyone else shall consider in determining whether a use is "fair" and thus non-infringing [7]. Those four factors are:

1. the purpose and character of the use, including whether such use if of commercial nature or is for nonprofit educational purposes
2. the nature of the copyrighted work
3. the amount and substantiality of the portion used in relation to the copyrighted work as a whole
4. the effect of the use upon the potential market or, or value of, the copyrighted work [6].

The U.S. Congress left the fair use guidelines vague. While both houses suggested guidelines in congressional reports that accompanied the legislation, these guidelines were not legally binding. Instead Congress deliberately delegated responsibility for determining fair use to the to the U.S. courts. Fair use lawsuits have produced many court rulings that have helped in defining fair use, for instance: delineating the differences between transformative vs. derivative as based on how the copyrighted

material is used. Other court rulings have pointedly linked the four factors together, though often one factor may be more influential on determining infringement. However much interpretation and application of the factors remains on a case by case basis. As a result there is much room for interpretation and individual judgment. According to a study that "investigated the knowledge, attitudes and experiences of media literacy educators regarding copyright and fair use, educators today have no consensus around what constitutes acceptable fair practices" [10]. Consequently, each U.S. institution has set up copyright policies and compliance procedures based on its own understanding of fair use. Disagreements between copyright holders and users are unavoidable, leading to more lawsuits.

U.S. Copyright law grants general infringement exemption to educational institutions under fair use. DMCA also included exemptions from criminal charges for higher education institutions. As a result educational bodies have avoided being overwhelmed with lawsuits. Currently, there are two notable lawsuits for fair use of digital works in progress involving Georgia State University and University of California, Los Angeles. In 2008 four academic publishers sued Georgia State University for putting 6,700 electronic journal articles on e-reserve in Blackboard or library system which the publishers asserted eroded into their revenues (Cambridge University Press, et al. v. Patton et al.). Since then Georgia State University has tightened its digital course reserve policies, but the courts has been ruling in favor of fair use by the university. After two years the battle still goes on in court. In 2010 University of California, Los Angeles was sued by the Association of Information Media and Equipment for putting streamed videos online for class use (AIME vs. UCLA). UCLA believed that it is fair use and would not back off their practice. Plaintiffs in both lawsuits claim that they are aware of copyright infringing practice under the pretence of fair use by many colleges and universities.

3 Digital Copyright in Selected European Countries

Copyright protections in Europe have a long history, dating back to the 19th century. These protections have gone through numerous revisions and changes continue as the European Union and individual nations confront new types of material such as internet publishing and digitization. Individual members of the European Union are responsible for formulating their own legal codes regarding copyright and keep them in agreement with the directives passed by the European Union, and statutes agreed upon in international treaties. As of February 2010 the European Union and its member states have no explicit policy regarding copyright for digitized materials. Several countries have attempted to address the matter themselves and encourage the Union to discuss and reach an agreement on a new directive for digital use.

In the last ten years the European Union has issued several Directives to enforce the treaties and better harmonize national approaches to copyright. The 2001 Directive (often referred to as the European Union Copyright Directive or EUCD), covered the nature of intellectual property rights, the Directive of 2004 addressed the enforcement of such rights, and the Directive of 2009 concerned legal protection for computer programs. All three agreements permit national legislation to provide for limitations of an author's copyright in certain special cases that do not conflict with a

"normal exploitation or the work" or the "legitimate interests" of the author. Each agreement specifically recommends copyright exceptions for educational and teaching purposes. Other optional exceptions as detailed in Article 5 of the EUCD including private copying, use of copyrighted material by libraries, museums and archives, ephemeral recordings, illustrations for teaching or scientific research, use for the benefit of people with a disability, press privileges, use for the purpose of quotations, caricature, parody and pastiche, use of political speeches and public lectures, use during religious or official celebrations, use of architectural works located permanently in public places, incidental inclusions of a work in other material, and use for the purpose of advertising the public exhibition or sale of artistic works. The list gave member nations great leeway by also allowing exceptions for additional cases of use having minor importance [11,12,13].

The EUCD exception list exemplifies the European approach to what is usually called "fair dealing" in member nations. Under this concept the moral right of the author stands as the center of copyright policy which serves to protect the property rights of the author concerning their creations. These perspectives empower legislators with a moral obligation to safeguard rights in a broad fashion that allows authors the opportunity to profit from the use of their creation while barring others from exploiting these creations. EU members have adapted these exceptions into their national laws in accordance with their own norms and traditions. Following the pattern of the Directive, member nations have specified specific limits or exceptions in long, detailed lists. This sets them apart from the American notion of fair use which outlines general criteria limiting copyright. The American method relies on courts to decide legitimate exceptions. The European approach emphasizes a legislative determination of valid exceptions [5,14].

Significantly, no EU Directive or member nation law explicitly addresses the digital age and problems confronting educational institutions regarding use of data streaming technology or web-based material. This has concerned many member nations, especially the United Kingdom and France who have initiated internal discussions to revise their copyright laws and bring them more in line with the digital era [15]. Their discussions have led the European Union to open deliberations regarding copyright for the digital environment [16]. The approach of the three leading European economic powers, Germany, France, and the United Kingdom have strongly influenced past EU directives on copyright and will likely play a large role in whatever new Directive is created concerning the internet, digital works, ISPs and enforcing copyright statutes. In line with European traditions, each of the three nations prioritize the moral rights of authors and reserves the right of integrity and attribution to the author, while recognizing the need for necessary exceptions to these rights for education and news purposes.

All three countries have incorporated the Berne and WIPO treaties as well as all EU directives into their copyright codes, and created detailed copyright exception lists. However each nation varies in their interpretation of treaty statutes. Differences in copyright exceptions, length of copyright protection, and the notion of "fair dealing" show why the European Union needs to continue its work at creating greater harmony among the member nations approaches to copyright.

Duration of copyright is one area not yet harmonized. The United Kingdom has the greatest variation in the length of copyright issued for different material. For literary,

dramatic, musical or artistic works, and film copyright is granted for the lifetime of the author or authors plus an additional 70 years following their death. A 70 year copyright is also given to anonymous works following their initial publication. Sound recordings and broadcasts copyright last for 50 years, while copyright for typographical arrangements or anthologies of previously published material lasts 25 years [17].

In Germany copyright expires 70 years after an author's death for general works, anonymous or pseudonymous works, cinematographic creations, and computer programs. The duration of Photographs and performances copyright is 50 years. Works published posthumously are given copyright for 25 years. Editions of non-copyright scientific works are given a copyright of 25 years [18].

France has the least complex length of copyright. In France all copyrights last 70 years after the death of the author or after a pseudonymous work is published. However, copyright is extended another 25 years if the author of a pseudonymous work is identified after the original copyright has expired [19].

As with the duration of copyright, approaches to "fair dealing" and copyright exceptions for education and research vary in each country. Of the three countries discussed in this paper only the United Kingdom gives a list of copyright exceptions under a general concept of "fair dealing." France and Germany have not codified any concept of "fair dealing." Therefore it is not surprising that the list of exceptions varies in each nation, though all three meet the exception statutes of Berne and the two WIPO treaties, as well as the three European Union directives.

The United Kingdom copyright code gives a far more extensive detailed list of what copyright exceptions are allowed than do Germany or France. Under the United Kingdom's "fair dealing" section exceptions include reproducing literary, dramatic, musical or artistic work for non-commercial research or private study provided that it is accompanied by a sufficient acknowledgement, and only 1% of the work is reproduced each quarter. A work can also be used for review or criticism or news reporting provided that work has previously been made available to the public. This exception does not apply to sound recordings, films or broadcast, photographs or unpublished material. Other educational exceptions in the United Kingdom include activities such as copying a literary, dramatic, musical or artistic work in the course of teaching as long as a reprographic process is not used (reprographic process means using a fax machine, photocopier or any appliance which makes multiple copies). "Fair dealing" also includes performing, playing or showing copyrighted works in a school, university or other educational establishment for educational purposes provided the audience consists of only teachers, pupils and other school officials, and recording a broadcast for non-commercial educational purposes in an educational establishment. Libraries are permitted to send copies of works to other libraries or for private use provided such use is for non-commercial purposes. Archives are permitted to copy a work for preservation or replacement purposes provided it is not reasonably practicable to purchase a copy of the item for that purpose [17].

Under its 2008 Law on Copyright and Related Rights Germany allows for the reproduction and distribution of limited parts of copyright material including works of language, photographs, and musical works for use or in education and research with the proper acknowledgement of the author with no definition of "limit." Teacher training schools are allowed to make individual copies of works which are included in

a school broadcast by recording the works on a video or audio medium provided the recordings are used only for instructional purposes and destroyed by the end of the school year unless the school pays the proper fee to the copyright owner. Museums, libraries and archives are allowed to digitize their collections and display them at electronic workstations with the stipulation that simultaneous access to each work is limited to the number of copies of the work owned by the library. Libraries are also permitted to copy and disseminate individual articles of newspapers and magazines and small portions of published works by mail, fax and other electronic form provided such copies are for non-commercial teaching or scientific research purposes. Archives are able to reproduce copies of a work if the archivist owns the copied material [18].

France permits short quotations or excerpts of copyright material to be used for critical, polemic, educational, scientific or informative uses. Works such as sheet music, and digital editions of published works can be reproduced for non-commercial educational purposes provided that the users consist of students, students, teachers or researchers directly concerned. Libraries are permitted to loan copyrighted material, but public or higher education libraries must pay a flat fee to a collection society for this exemption. The fee is based on a library's number of registered users. The collection society then pays remunerations to the publisher and copyright owner. Libraries, museums, and archives are permitted to reproduce works on dedicated terminals for non-economic use with no restriction regarding simultaneous use. Archives can reproduce a work for conservation or preservation purposes [19].

In all three nations, it is the non-commercial or economic use of the material which permits all acts of "fair dealing." The examples of exceptions listed in the previous paragraphs are given to illustrate the ways each nation has approached codifying similar educational uses of copyright material. They are not meant to be a complete list of all such exceptions. France has the fewest number of exceptions codified, while the United Kingdom has the most. Of the exceptions described, it is notable that the United Kingdom has not addressed the networking of copyright material on dedicated library terminals as has Germany and France.

To address repeated copyright violations by internet "pirates" and meet the EU's 2009 directive, the United Kingdom and France have adopted a so-called "Three-Strike" rule aimed at Internet Service Providers. This approach has been very controversial and has led both nations to urge the European Union to revisit copyright laws in order to better address the digital age. The United Kingdom has already initiated internal discussions about copyright law. Both the United Kingdom's Digital Economy Act of 2010 and France's passing of its Haute Autorite Pour la Diffusion des Oeuvres et la Protection des Droits sur Internet (Hadopi} law have pressed the EU to open new copyright deliberations in 2011 [20,21,22].

Germany was the last EU member nation to meet the European Union's 2004 directive. Its "second basket" revisions of 2008 addressed several digital age issues such as copying or downloading non-compliant material illegally posted on the Internet. The new law also permitted libraries to reproduce works and distribute or transmit the copies by e-mail. Germany has also begun discussions on a "third basket" revision to more thoroughly meet the copyright needs of the modern "online" era [23,24].

4 Comparison

While both European Union members and the United States are committed to find a balance between the rights of the copyright holder and the public interest, their approaches to achieving this balance continue to reflect the differing cultures and societal outlooks of each nation. The U.S. stood apart from Europe until its extensive rewriting of copyright law in 1976. Even then, it was not until 1988 following additional revisions to its copyright law that the United States was able to formally join the Berne Convention agreements. Since then the U.S. has been a part of every major international agreement regarding copyright signed by the European Union members and most of the other industrial nations of the world. All the copyright agreements require copyright notice (digital management information) to be attached to digital as well as printed works. This has created greater worldwide harmony regarding copyright rights, including recognition of international copyright law.

Prior to its passage of the 1976 Copyright Act, the United States rooted its copyright law upon the date of publication or broadcast of a work. In the 1976 Act the United States adopted the European tradition of basing copyright on the life of the author. Both the United States and Europe now recognize the general exclusive rights of copyright holders as the creator's life time plus 70 years after the death, though differences still remain for other types of material such as anonymous works, anthologies, and some digital formats.

Recognition of international copyright laws has not solved all conflicts between the copyright laws of individual nations, especially regarding internet or digitized material. Though all of the nations reviewed in this paper recognize general educational and library exemptions for the use of copyrighted material, the variation among allowable exemptions create issues regarding copyright infringement and enforcement. While each nation allows libraries to participate in interlibrary loan programs, copy material for patron use, and allow access to scholarly database aggregates, jurisdictional issues concerning copyright infringement have not been fully addressed. The differing notions of fair use and fair dealing may produce misinterpretation and misapplication of correct copyright law when dealing with the loaning or copying of material between libraries of Europe and the United States. Educators and librarians need to be aware of possible conflicts and the potential for unintended infringement, such as fee requirements, when obtaining material for researchers and educators.

In general the copyright law of the United States consists of broader, ambiguous principles while European statutes compile specific lists of author rights, exceptions, and enforcement of copyrights. The European detailed lists are very specific regarding an author's rights and certain restrictions to those rights, especially for education. Overall, the United States seems to emphasize a more public-friendly approach to the use of copyrighted material than do European countries. Europe in general seems to place greater emphasis on the protecting the property or moral rights of authors. This is especially true for the provisions regarding "fair use" or "fair dealing." As we have previously noted, the U.S. Congress deliberately left much of the decision-making regarding legal "fair use" of copyright material to its courts. Obtaining final determination of copyright issues can take years as the case works it way through the various levels of appeal. The European approach can also be delayed or derailed in

addressing new issues that arise in the digital age due to legislative debate and disagreement. Neither approach has been fully successful or satisfying in addressing the educational use of copyright material in the digital age.

5 Conclusion

The arrival of the digital age has brought new challenges to copyright that must be addressed. The variety of digital works continues to grow. The array of formats such as video, Flash, DVD, CD, streaming media, digital texts, MP3, I-Pod, digital images, email, net communication, Web, Flash tutorials, etc, seems endless. The Internet's growth as a channel of transmitting digital data in turn continues to give rise to new and unique situations. It is hard for digital copyright law to keep up with each new scenario. The methods used by the United States and Europe to address these ongoing developments have had varying degrees of success. It is appropriate to ask which approach seems more appropriate.

The United States legislature has only lain down the principles for fair use. It has not codified a definition or guidelines for infringement exemptions. Instead it has been left to the U.S. judicial system to serve as the major force in defining fair use in digital works in the United States. The result has been an uneven interpretation of the four fair use factors leading to many education leaders to either hold back on allowing copyrighted material to be used in classrooms or they have allowed teachers to use all available materials in the hope that, if sued, the courts will rule it permitted under fair use.

The list of fair dealing exemptions in the three European nations may not be as vague as U.S. law, but still does not offer educators a good deal of guidance regarding the amount of copyrighted material permitted to be used in education. Only the United Kingdom has codified a 1% amount. Germany grants a "limited" amount, while France permits "brief" excerpts. All three permit libraries to reproduce material for research and use by other libraries via interlibrary loan.

The WIPO agreements were important first steps in addressing copyright in the digital age, as have the European Union Directives and the 2007 Rome II agreement. In the United States, the Digital Millennium Copyright Act (DMCA), the Digital Choice and Freedom Act and the TEACH act give the U.S. the lead in updating its copyright laws and addressing the educational needs of the digital age. One important DMCA provision with important consequences for fair use and education requires the Librarian of Congress to review the digital law every three years and grant exemptions to copyright infringement subject to renewal at the next review. European nations are just starting to discuss these issues and their first steps have consisted of addressing copying illegal material from the Web through legislation such as France's Hadopi act or the United Kingdom's Digital Economy Act for internet service providers. It remains to be seen if this approach is successful. Certainly more need to be done.

Our research into copyright law has revealed the need for negotiated compromises between the EU and U.S. One important first step would be a more fully delineated the concept of fair use or fair dealing with appropriate codified guidelines. The United States rather than relying on its courts to determine fair use should incorporate some version of the 1976 congressional reports guidelines into its legal code regarding the nature and

amount of a work that can be copied or used under fair use. Likewise, the members of the European Union should work at establishing clearer guidelines of their own under their concept of fair dealing. These guidelines should be negotiated on an international basis and thus achieve a more uniform fair use/dealing concept and law worldwide. The guidelines should recognize the educational needs for accessing copyrighted material while preserving the moral rights of the authors.

Another problematic area of copyright that needs to be addressed is educating teachers about their classroom rights under copyright law. A recent U.S. study interviewed sixty three educators who taught media literary both in higher education and k-12, The teachers had conflicting ideas about digital copyright [10] and its classroom application.The lack of understanding and consensus about copyright law led to two scenarios: either educators become overly restrictive in using copyrighted works in teaching or they totally disregard and may violate copyright law. When the first case occurs, faculty tends to avoid using copyrighted materials in teaching out of fear for infringement, depriving themselves of the benefits from education exemptions and fair use are not being. In the latter case, faculty members use copyrighted materials freely regardless of copyright law. Such recklessness may encounter lawsuits. In either case, educators do not take full advantage of fair use. To deal with this widespread problem, there should be an organized effort at all levels of education to enlighten faculty about their digital copyright rights.

Digital copyright relating to fair use is an interesting and complex subject to discuss. Both the United States and Europe face severe challenges. Copyright is not only an important incentive for research and economic development; it is also an important protector of artistic, non-commercial expression and educational advancement. The goal of the EU, U.S. and other nations should be agreements that incorporate a balance of copyright specificity while remaining open to new developments that lie ahead.

References

1. Lyons, M.G.: Open Access Is Almost Here: Navigating through Copyright, Fair Use, and the TEACH Act. J. Cont. Ed. 41(2), 57–65 (2010)
2. Berne Convention, http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html
3. World Intellectual Property Organization: WIPO Copyright Treaty, http://www.wipo.int/treaties/en/ip/wct/pdf/trtdocs_wo033.pdf
4. World Intellectual Property Organization: WIPO Performances and Phonograms Treaty, http://www.wipo.int/treaties/en/ip/wppt/trtdocs_wo034.html
5. Lohmann, F.: Fair Use and Digital Rights Management: Preliminary Thoughts on the (Irreconcilable?) Tension between Them. Electronic Frontier Foundation, http://w2.eff.org/IP/DRM/fair_use_and_drm.html
6. United States Copyright Office: "Fair Use." Copyright: the U.S Copyright Office, http://www.copyright.gov/fls/fl102.html
7. Beebe, B.: An Empirical Study of U.S. Copyright Fair Use Opinions, 1978-2005. University of Pennsylvania Law Rev. 156(3), 549–624 (2008)
8. United States Copyright Office: The Digital Millennium Copyright Act of 1998: U.S. Copyright Office Summary, http://www.ictregulationtoolkit.org/en/Publication.1466.html

9. American Library Association: Distance Education and the TEACH Act, http://www.ala.org/Template.cfm?Section=distanceed&Template=/ContentManagement/ContentDisplay.cfm&ContentID=25939
10. Hobbs, R., Jaszi, P., Aufderheide, P.: The Cost of CopyrightConfusion for Media Literacy. Online Submission. ERIC ED499465 (2007)
11. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society. Official J. of the E.C (2001)
12. Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the Enforcement of Intellectual property rights. Official J. of the E.U (2004)
13. Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs. Official J. of the E.U (2009)
14. Senftleben, M.: Bridging the Differences between Copyright's Legal Traditions–the Emerging Ec Fair Use Doctrine. J. Copyright Soc. U.S.A 57, 521–552 (2010)
15. Korba, J.: United Kingdom Reviews Copyright Laws. Intellectual Property Brief (November 9, 2010), http://www.ipbrief.net/2010/11/09/united-kingdom-reviews-copyright-laws/
16. Jowitt, T.: European Parliament Calls for EU-Wide Copyright Law. eWeek (September 23, 2010), http://www.eweekeurope.co.uk/news/european-parliament-calls-for-eu-wide-copyright-law-10010
17. Office of Public Sector Information of the National Archives of the UK: Copyright Designs and Patent Act (1988), http://www.legislation.gov.uk/ukpga/1988/48/contents
18. Bundesministerium der Justiz: Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz), http://www.gesetze-im-internet.de/urhg/BJNR012730965.html
19. Legifrance.gouv.fr: Code de la propriété intellectuelle: Version consolidée au (mars 16, 2011), http://www.legifrance.gouv.fr/affichCode.do;jsessionid=98C01F4A7DAA823CA16E428FC087DBB4.tpdjo05v_1?idSectionTA=LEGISCTA000006161636&cidTexte=LEGITEXT000006069414&dateTexte=20110320
20. Office of Public Sector Information of the National Archives of the UK: Digital Economy Act (2010), http://www.legislation.gov.uk/ukpga/2010/24
21. BBC News: UK Copyright Laws to Be Reviewed, Announces Cameron (November 4, 2010), http://www.bbc.co.uk/news/uk-politics-11695416
22. Schöpfel, J.: The New French Law on Author's Rights and Related Rights in the Information Society, http://www.deepdyve.com/lp/emerald-publishing/the-new-french-law-on-author-s-rights-and-related-rights-in-the-fkbIbqvtAs
23. Weisser, R.: "Second Basket" of German Copyright Reform Closes Loophole on File-Sharing, Free Treasures from Studios' Archives, http://www.mwe.com/index.cfm/fuseaction/publications.nldetail/object_id/db6e5714-77e4-4738-8293-ca37ffaa36ed.cfm
24. STM: Update "Second Basket" German Copyright Law, http://www.stm-assoc.org/2007_10_01_German_Copyright_Law_Update_Second_Basket.pdf

INVISQUE: Technology and Methodologies for Interactive Information Visualization and Analytics in Large Library Collections

B.L. William Wong, Sharmin (Tinni) Choudhury, Chris Rooney, Raymond Chen, and Kai Xu

Interaction Design Center, School of Engineering and Information Sciences, Middlesex University, Hendon, London, NW4 4BT England
{w.wong,t.choudhury,c.rooney,r.chen,k.xu}@mdx.ac.uk

Abstract. When a user knows exactly what they are looking for most library systems are adequate for their needs. However, when the user's information needs are ill-defined - traditional library systems prove inadequate. This is because traditional library systems are not designed to support sense making rather for information retrieval. Visual analytics is the science of analytical reasoning facilitated by interactive visualizations and visual analytics systems can support both sense making and information retrieval. In this paper, we present INVISQUE – an approach and experimental software for interactive visual search and query. INVISQUE uses an index card metaphor to display library content, organized in a way that visually integrates attributes such citations and date published, making it easy to pick out the most recent and most cited paper. It uses design techniques such as focus+context to reveal relationships between documents, while avoiding the "what-was-I-looking-for?" problem.

Keywords: Visual Analytics, Information Visualization, User Interface, Interactive Visualization.

1 Introduction

Visual analytics is the science of analytical reasoning facilitated by interactive visualizations [1]. Visual analytics combines automated analysis techniques with interactive visualizations of large and complex data sets for supporting understanding, reasoning and decision making [2]. Navigating the large collections of most digital libraries can be a very complex task, especially when approaching such collections with an ill-defined information need.

In contrast, well-defined information need is knowing exactly what you want, e.g. I want a book on visual analytics titled "Illuminating the Path"; whereas an ill-defined information need is only having a vague idea of what you are searching for, e.g. I want information on visual analytics but I am not sure what is out there and what will be useful to me. When the information need is well-defined, information-seeking becomes a simple information retrieval (IR) task [3]. However, when the information

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 227–235, 2011.

needs are for more complex mental activities such as learning and decision making, IR is necessary but not sufficient [3]. Existing library systems are well adapted at IR but they are not adequate for the far more complex task of information exploration and sense-making that is necessary when a user's information need is ill-defined and they are learning as they are exploring the information to (1) define exactly what their information needs are, (2) before they do the actual retrieval.

The JISC-funded User Behaviour in Resource Discovery (UBiRD) [4] study investigated why expensive electronic library resources where underutilized. The study found that many electronic resource discovery systems currently used within libraries (e.g. Emerald, ISI, etc) distract users from focusing on the content, analysis and evaluation that would help them learn and make sense of the resources discovered [4]. A visual analytics capable digital library system, such as the one proposed in this paper, has the potential to allow users to focus on content, analysis and evaluation and therefore, aid them to learn and make sense of what they discovered.

INVISQUE [5], Interactive Visual Search and Query Environment, developed based on findings of the UBiRD study, is an approach and experimental software for interactive visual search and query that can function as a visual analytics library system that encourages users to focus on content, analysis and evaluation. INVISQUE uses an index card metaphor for displaying digital library content. The INVISQUE visualization, which employs index-card visualisation, attempts to assist user by supporting in-context viewing by clustering index-cards by subject and by mapping information against visually integrated dimensions by ordering the index-cards on both the X and Y axes. In Figure 1, the cards are organized according to the 1st Author of the paper on the X axis and the year published on the Y axis.

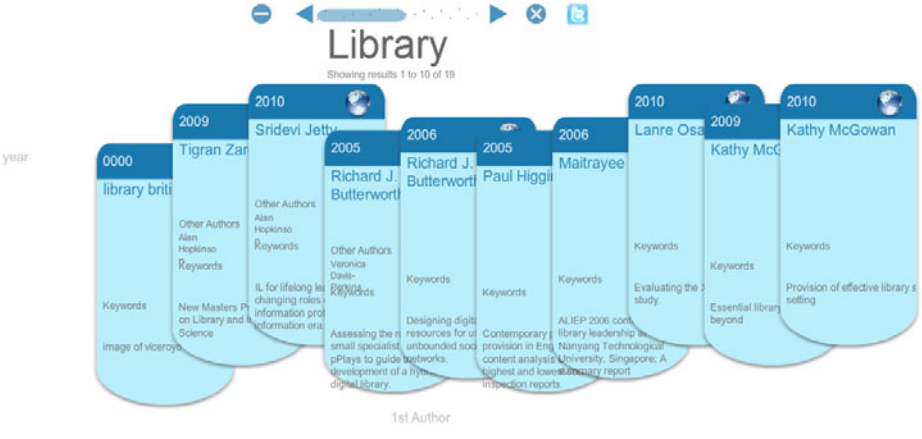


Fig. 1. INVISQUE Interface

INVISQUE also uses information design techniques such as focus+context to present detailed information while retaining the context of the search within the visual field of view. This helps users maintain their orientation within various searches, minimize their chance of getting lost in the data and experiencing the "what-was-I-looking-for?" or 'WWILF-ing' problem, which is a problem with the current search

systems [6]. INVISQUE also attempts to support higher order functions such as finding semantically meaningful relationships between data entities by revealing resources sharing the same or semantically similar relationships (Fig. 3 and 4), when prompted to do so [7].

In this paper, we first discuss the background and motivations behind INVISQUE before illustrating the functional capabilities of INVISQUE using a use case scenario. We then discuss the system structure and implementation of INVISQUE before presenting our planned future work and concluding.

2 Background and Motivation

INVISQUE was developed as a possible solution to the problems identified in the UBiRD study: poor usability, high complexity, and lack of integration in many electronic resource discovery systems, acting as a barrier to information search and retrieval[4]. The study also found that users did not understand how to assess the quality of materials they found, which the UBiRD researchers attributed to poor information literacy [4]. However, Spink offers a different view and suggests that the user's ability to rank relevance and irrelevance of information is based on 1) information problem definition, 2) search intermediaries' perceptions that a user's question and information problem has changed during the mediated search interaction, 3) personal knowledge due to the search interaction, and 4) criteria for making relevance judgments [8], i.e. the more well-defined users' needs are, the more they are able to judge relevance. By contrast, the more ill-defined the users' needs are the more they need guidance in order to judge relevance. The study also observed that the level of the user's domain knowledge alters their behavior [4].

Both Spink and the UBiRD study observed a process of progressive understanding from the perspective of the user where the user often starts with an ill-defined search criterion in their mind but continuously refines the criteria based on the results they obtain from IR tools [4,9,10]. In addition, studies conducted by Kodagoda on the search behavior of low and high literacy users, showed that low and high literacy users demonstrated markedly different search behaviors with low literacy being at a disadvantage with conventional list form result displays [11]. Motivated by these studies, INVISQUE was developed as a potential solution to the issues identified [5].

Combining multidimensional visualisations and dynamic queries is not a new concept. Ahlberg and Shneiderman visualized search results using two dimensional scattergrams and provided sliders to filter the data [12,13]. HomeFinder used dynamic queries and sliders for user to control visualization of multidimensional data [14]. More recently Stasko et al., developed the JIGSAW system that provides multiple coordinated views of document entities emphasizing visual connections between entities across the different documents [15]. INVISQUE combines the information visualization concepts mentioned above with a modern visual interface and emerging interaction technologies with the goal being to assist users the sensemaking process.

INVISQUE is heavily influenced by the Pirolli & Card model of intelligence sensemaking, shown in Figure 2, [16]. The Pirolli & Card model came out of analysis of sensemaking activities that took place amongst intelligence analyst [16]. However, the model can be applied to sensemaking in other domains, such as the digital library domain.

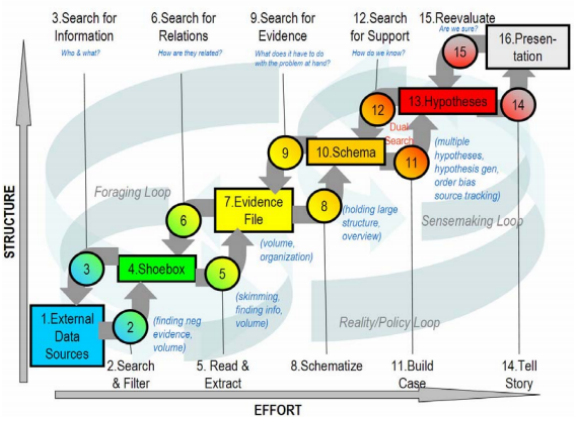


Fig. 2. Pirolli & Card Model Of Sense-making [16]

The INVISQUE system is supported by a hybrid adaptive architecture [17]. An adaptive architecture changes its structure based on use and demand [18]. Advocates of the architecture propose adaptive system based on adaptive software architecture to be the key to achieving the goal of retaining full application plasticity throughout the software's lifecycle and that are as easy to modify on the field as they are on the drawing board [18]. We aim to take advantage of adaptive architectures to make INVISQUE more robust, scalable and changeable.

3 INVISQUE Interface – An Use Case Demonstration

The best way to highlight the properties of INVISQUE is through a use case. Our INVISQUE demonstrator has been connected to a range of different data sets such as the National Counter Terrorism Center's (NCTC) Worldwide Incident Tracking System (WITS) dataset, Google cinemas and e-gov social services datasets. For this paper, our use case describes INVISQUE connected to the Middlesex University ePrints Repository and the typical information search of an early-stage PhD. Firstly, while the PhD student will become an expert of sorts in their chosen field by the end of their PhD candidature, at the beginning of said PhD candidature – the student is a novice with their very topic of their PhD changing and evolving and thus the student's information needs are highly ill-defined.

Secondly, the PhD student would fall in the category of people who, according to Spink and the findings of UBiRD [4,8] would not necessarily know what is relevant and what is irrelevant, adding to the challenges in ill-defined nature of the students query. Lastly, as defined by Marchionini, the PhD student is engaged in complex mental activities of learning and decision making while exploring the literature [3].

Let us assume that this student is doing their PhD in interaction design under the supervision of Wong. Very likely, the first search terms they would enter into an IR tool is "wong" and "design". Figure 3 shows the two clusters of index cards that are produced from these two searches. By having both search results in the same visual field of view, the student immediately engages in sensemaking by explore

relationships through the focus+context features of INVISQUE. As shown in Figure 3, when an index card is selected – other cards which share traits with the highlighted card come into focus, while non-related cards fade into the background. This gives the students information about relationships that they would not easily get from a list display, and assists in the sense-making process by highlighting common traits such as shared keywords, authors and publication years. In Figure 3, the common trait being highlighted across the two clusters is the author.

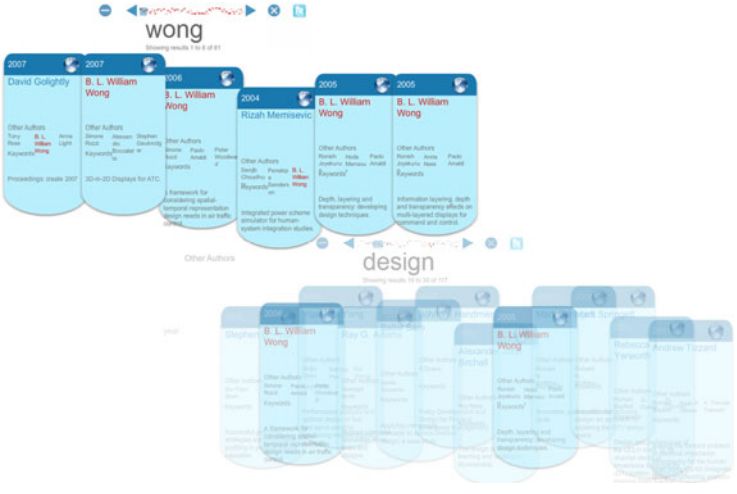


Fig. 3. Display Showing Two Result Clusters With Focus+Context

Now armed with more information the student can launch a third search based on, for example, keyword discovered amongst the first set of results. Alternatively, the student can start "shoeboxing", see Figure 2, gathering and categorizing publications they are interested. INVISQUE also allows the gathering of selected cards into a new cluster. The student can also customize the X-Y axes to specify the manner in which the index cards are arranged or zoom into the *Contextual Interval Slider*, shown in Figure 4, which provides the student an indication of how many other records within the dataset that share the same traits. In addition, the student also has the ability to drill down and bring-up addition information about a selected index card – for example, the PDF of a journal article.

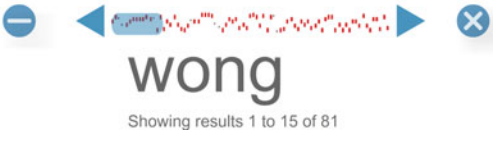


Fig. 4. Records Of Interest Highlighted In The Contextual Interval Slider

INVISQUE embodies the concept of an infinite canvas which allows for a potentially limitless amount of search clusters to be displayed together. The infinite canvas concept has the potential of being adapted and deployed into collaborative environments, enabling multiple users to search simultaneously.

4 INVISQUE System Architecture

INVISQUE is very much a working prototype with new features and functionality being added to the system almost on a daily basis. It is also not just a digital libraries visual analytics system. We have put INVISQUE to other uses, such as its ability to act as an information kiosk for entertainment information [19]. The entertainment information kiosk version of INVISQUE differs mainly in the interface, with more emphasis being placed on multi-touch based user interactions as opposed to the point and click mouse-based interaction emphasized in the digital library version of INVISQUE. However, while the interface differs significantly between deployments of INVISQUE – the overall system architecture is designed to remain consistent. Indeed, one of the reasons we opted for an adaptive architecture is so that INVISQUE can change depending on the use to which it is placed.

In Section 2, we called the INVISQUE architecture a hybrid adaptive architecture. The hybridization is with a 3-tier client-server architecture which forms the primary structure of the INVISQUE architecture. In the "middle-tier" of the n-tier architecture, we have imbedded a rule-based *Architecture Controller* that orchestrates functional components based on data-type, data-load and functional prompts from the interface. The architecture is illustrated in Figure 5.

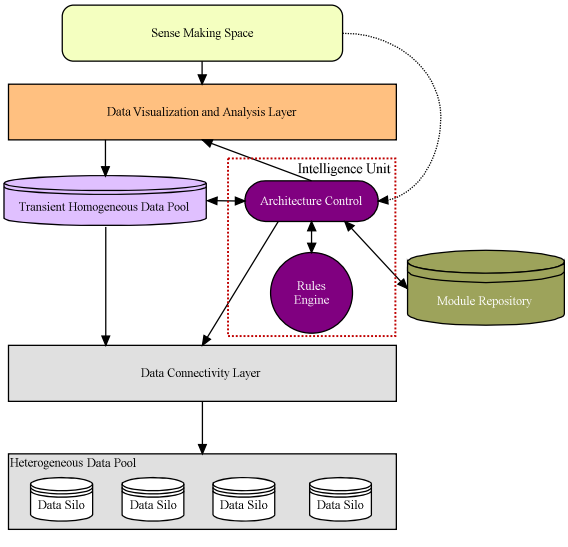


Fig. 5. INVISQUE Architecture

To explain the architecture, at the data-tier level we have a number of silos containing heterogeneous data. The UBiRD study revealed that users prefer integrated systems [4] and INVISQUE does aim to provide an integrated view of multiple data sources. Therefore, the first layer of functionality in the middle-tier is the *Data Retrieval Component Orchestration Layer*. The exact functionalities are deliberately abstracted and will be determined by the data silos the system has to access.

For example, the Middlesex University ePrints Repository to which INVISQUE is currently connected to is a MySQL database. However, we are also working on a project that connects INVISQUE to data that is in XML format. INVISQUE will work with both the MySQL database and the XML database in tandem and display a merged result set when a keyword search is executed. This is where the *Transient Homogenous Result Pool* comes in. The idea is that results from underlying data silos will be gathered in the pool and made available to the upper tier functions; so that the upper tier do not have to consider data structure. Additionally, this separation also means that the *Architecture Controller* has greater control over adaptation.

Once the results are in a collected format, the INVISQUE interface is free to execute operations on the result pool. Again, the *Data Visualization and Analysis Component Orchestration Layer* is kept deliberately abstracted because the idea is to add and remove functionality quickly and easily as required, thus scaling and changing the system dynamically. This is because INVISQUE is only at a fraction of what it is capable of. So that we can easily add and remove functionality and adapt the same system for multiple uses, the functional layer has to be as abstract as the data retrieval layer with the nucleus of the architecture, the *Architecture Controller*, pulling from a pool of components the required components and orchestrating them on the relevant layer based on predefined rules that are triggered by both interactions happening in the user interface and changes in the data-tier.

5 Evaluation and Summary

At the time of writing, as part of her PhD, Kodagoda had evaluated INVISQUE on high and low literacy users in the context of finding on-line social service information [6]. Kodagoda tested the hypothesis that the context layering offered by INVISQUE would reduce premature search abandonment when compared with traditional hierarchical website layout. The results to date are in favor of INVISQUE [6]. In the near future we will be conducting scalability tests and re-design once we complete the integration of a 2 terabyte journal archive provided to us by a major journal publisher, as mentioned in Section 4.

INVISQUE aims to present the design for the next generation of information search and retrieval systems that would support semantic analysis and access to massively large data sets. The design we have developed shows how we can visualize and present search results and we can facilitate interaction with the visualization.

We have moved away from the conventional list-style arrangement, and instead represent information by the use of index cards in a 2-dimentional space. The open canvas allows users to perform multiple searches while keeping the context of the complete search space.

We believe that INVISQUE's visibly novel user interface, backed-up with its hybrid adaptive architecture can overcome the shortcoming of existing library systems.

Acknowledgments. INVISQUE was originally funded by the JISC Rapid Innovation program, Grant Ref. Num. IEDEVC19/RI. We would also like thank Middlesex University for its continued support as well as the Interaction Design Center team who have had an input into INVISQUE.

References

[1] Thomas, J.J., Cook, K.A.: Illuminating the Path: The research and development agenda for visual analytics. IEEE Computer Society, Los Alamitos (2005)
[2] Kei, D., Andrienko, G., Fekete, J.-D., Gorg, C., Kohlhammer, J., Melancon, G.: Visual Analytics: Definition, Process, and Challenge (2008)
[3] Marchionini, G., White, R.W.: Information-seeking support systems. IEEE Computer 42, 30–32 (2009)
[4] Wong, W., Stelmaszewska, H., Bhimani, N., Barn, S., Barn, B.: User Behaviour in Resource Discovery: Final Report (2009)
[5] Khan, N.: INVISQUE, INteractive VIsual Search and QUery Environment (2010)
[6] Kodagoda, N.: PhD Thesis of Neesha Kodagoda. Middlesex University (2011)
[7] Stelmaszewska, H., Wong, B.L.W., Attfield, S., Chen, R.: Electronic resource discovery systems: from user behaviour to design. In: Proceedings of the 6th Nordic Conference on HumanComputer Interaction Extending Boundaries, pp. 483–492. ACM, New York (2010)
[8] Spink, A.H., Greisdorf, H., Bateman, J.: From highly relevant to not relevant: examining different regions of relevance. Information Processing and Management 34, 599–622 (1998)
[9] Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology 52, 226–234 (2001)
[10] Spink, A., Saracevic, T.: Human-computer interaction in information retrieval: nature and manifestations of feedback. Interacting with Computers 10, 249–267 (1998)
[11] Kodagoda, N., Wong, W.B.L., Khan, N.: Information seeking behaviour model as a theoretical lens: high and low literate users behaviour process analysed. In: ECCE, pp. 117–124 (2010)
[12] Ahlberg, C., Shneiderman, B.: Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. In: ACM CHI Conference on Human Factors in Computing Systems, Boston, USA, pp. 313–321 (1994)
[13] Shneiderman, B., Ahlberg, C.: AlphaSlider: A compact and rapid selector. In: ACM CHI Conference on Human Factors in Computing Systems (1994)
[14] Williamson, C., Shneiderman, B.: The Dynamic HomeFinder: Evaluating dynamic queries in a realestate information exploration system. In: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1992)
[15] Stasko, J., Görg, C., Spence, R.: Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization 7, 118–132 (2008)

[16] Pirolli, P., Card, S.: The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In: Proceedings of the 2005 International Conference on Intelligence Analysis (2005)
[17] Choudhury, S.T.: Loculus: An ontology-based information management framework for the Motion Picture Industry. Queensland University of Technology (2010)
[18] Oreizy, P., Gorlick, M.M., Taylor, R.N., Heimbigner, D., Johnson, G., Medvidovic, N., Quilici, A., Rosenblum, D.S., Wolf, A.L.: SELF - ADAPTIVE An Architecture-Based Approach to Self-Adaptive Software. IEEE Intelligent Systems, 54–62 (1999)
[19] Wong, W., Rooney, C., Chen, R., Xu, K., Kodagoda, N.: INVISQUE: Intuitive Information Exploration through Interactive Visualization. In: ACM CHI Conference on Human Factors in Computing Systems (2011)

An Evaluation of Thesaurus-Enhanced Visual Interfaces for Multilingual Digital Libraries

Ali Shiri[1], Stan Ruecker[2], Lindsay Doll[3], Matthew Bouchard[2], and Carlos Fiorentino[4]

[1] School of Library and Information Studies
[2] Humanities Computing Program
[3] Faculty of Education,
[4] Department of Art and Design
University of Alberta
{ashiri,sruecker,carlosf}@ualberta.ca,
matt.bouchard@gmail.com

Abstract. In this paper, we describe a comparative user evaluation of two multilingual thesaurus-enhanced visual user interfaces, namely T-Saurus and Searchling, developed for digital libraries. The study used 25 academic users carrying out three search tasks on both user interfaces to the UNESCO digital portal, holding 400,000 documents. It applied usability and affordance strength questionnaires, interviews, thinkalouds, and direct observation to investigate users' evaluation of the key components of both user interfaces, namely multilingual features and thesaurus and search functions. The empirical data gathered will be useful for designers of search interfaces that use thesaurus and multilingual features. Results of the study show that users were able to successfully carry out the search tasks using thesaurus-enhanced search interfaces. However, they preferred Searchling for its flexible language option, thesaurus browsing and visualization.

Keywords: Visual Interfaces, Multilingual Thesauri, Multilingual Digital Libraries, Information Retrieval, User Evaluation.

1 Introduction

Highly interactive and dynamic user interfaces for exploratory browsing and searching of digital information collections have been the focus of several recent studies. White et al. [1] note that in exploratory search, users generally combine querying and browsing strategies to foster learning and investigation. Marchionini [2] argues that semantically rich user interfaces have the potential to assist users in formulating queries, forming context for a particular search and exploring and gaining a comprehensive view of collections. Providing useful semantic assistance, particularly through visualization, within user interfaces of digital libraries requires research into the type of visualization and the associated features to support users in the exploration, searching and browsing of the collection. There have been a number of thesaurus-enhanced visual user interfaces that have been subject to evaluation. Déjà vu [3], uses the Library of Congress Thesaurus of Graphical Materials in its interface to provide a browsing facility for retrieval in a catalogue of digital media. A

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 236–243, 2011.

user evaluation of the interface has shown that the process of browsing through the thesaurus terms in Déjà vu improves users' understanding of the relationship between the archive materials and the cataloguing resources. Sutcliffe et al. [4] evaluated users' interaction with a thesaurus and results browser and found that better searchers used the visualizations more effectively and spent longer on the task, whereas poorer performances were attributable to poor motivation, difficulty in assessing article relevance and poor use of system visualizations. McKay et al. [5] evaluated thesaurus-enhanced search interfaces for digital libraries and found that participants used the thesaurus less frequently when it was in a separate window, and that the multiple independent window interfaces were awkward to use. They preferred that the thesaurus act semi-automatically (that is, that it suggested search terms) rather than it automatically inserts thesaurus terms into the search, or that it forced the user to manually search the thesaurus for terms of interest. Blocks et al. [6] found that a thesaurus-enhanced search interface was successful in allowing a person with little knowledge of the interface to make use of its functionality. However, the prototype interface did not provide non-expert searchers with sufficient guidance on query structure or when to use the thesaurus within the search process. In our own previous work, we [7] evaluated a Bilingual version of Searchling and found that integrating search and browsing features was particularly useful and that the semantically enhanced visual interface was most useful at the beginning of a research project on an unfamiliar topic, because users could start by browsing through general categories for relevant terms and the Thesaurus could help them narrow or broaden their search. Other researchers [8] have found that the provision of facets on the user interface affected users' search and browsing behaviour. Users expressed interest in the ways in which facets provided starting points for browsing and searching.

In this paper we report empirical evaluation of two visual, exploratory user interfaces that take advantage of dynamic views supported by the UNESCO multilingual thesaurus in English, French and Spanish languages. The key features of the two interfaces that we have developed are a) combining searching and browsing, b) supporting dynamic exploration of the conceptual structure of a thesaurus, c) providing dynamic term relation features to give high level overviews of the terms and the collection, d)) supporting multilingual search and retrieval within the UNESCO digital collections and e) utilizing a novel technique to implicitly show thesaural relationships using colour, size and distance. A comparative user evaluation of the two user interfaces was carried out to examine the multilingual and visual interface features and functionalities that support users in exploring semantic information, formulating queries and interacting with digital information. The results from the study contribute to our understanding of the factors affecting users' interaction with visual user interfaces that provide thesaural semantic support for query formulation and information exploration.

2 Visual User Interfaces

We developed two different visual user interfaces using the UNESCO multilingual thesaurus. They are called Searchling and T-Saurus, and their working prototypes are available at: http://thesaurusbrowser.info. The theoretical framework behind these

interfaces is reported in [7] and [9]. The Searchling user interface provides the user with the following three spaces within a single screen: the thesaurus space, the query space, and the document space (Figure 1). The Thesaurus space is on the left. It includes a browsable side panel of high-level categories, next to a list of thesaurus terms. Each term has a number beside it, which indicates how many documents in the collection contain the term. When a term is queried or clicked, it moves to the top of the list and all related terms from the thesaurus appear below it. The table to the right of the Thesaurus list indicates related terms that are broader, narrower, preferred or non-preferred compared with the selected term; the user can also sort by these categories. Finally, there is a language switch at the top of the Thesaurus list. The Query space is located in the right panel of the screen. Users can search for a single term in the thesaurus by entering it in the query box, choosing a language, and clicking the button labeled "Find in Thesaurus." If the term is entered in English but the user selects Spanish or French as the query language, Searchling will search for the corresponding Spanish or French term, but the English term will also always be visible as a microtext satellite below the query.

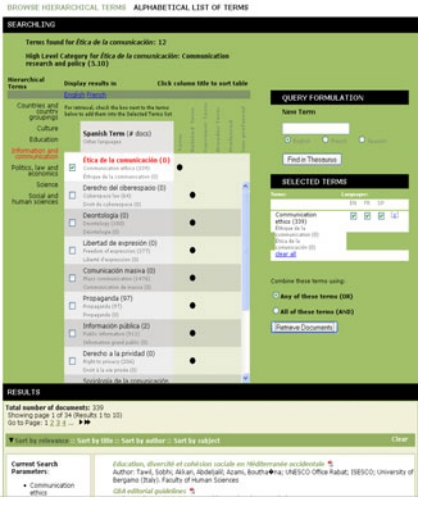


Fig. 1. Searchling interface

Fig. 2. T-Saurus interface

When users decide to add a term to their query, they do so by checking the box next to it in the thesaurus list, and it is added to the Selected Terms list on the lower right side panel. The Document space forms the third section of the screen, running across the bottom. Figure 2 shows the T-Saurus search user interface. The user interface makes use of visual objects, size, colour, location, zoom in and zoom out features to distinguish between various types of thesaurus terms and their relationships. Figure 2 shows a core of visual elements consisting of a set of "buckets" organized in the center of the screen. It shows the size of the buckets that represents the number of matches for a particular term, while proximity and opacity represent scope and accuracy of the term in relation to pre-established hierarchies for

the query: main term, related terms, more specific, more general and synonymous terms. The Query space is located across the top and on the right side of the screen while the Thesaurus space is located on the left and in the centre. Users can search for a single term in the thesaurus by entering it in the query box at the top of the page and clicking the Find button. If the term exists in the thesaurus it will appear in the centre of the screen with a number in parentheses beside it, which indicates the number of documents in the collection that include the selected term. Users can also browse all the terms in the thesaurus using the panel on the left, which can be sorted either alphabetically or hierarchically by category. When a term in the list is clicked, it will appear in the centre of the screen. When a term is selected by either method it is represented by a square in the central Thesaurus space. By utilizing the checkboxes in the bottom of the right-hand panel, users can choose to view the thesaurus terms that are related, narrower (more specific), broader (more general), and preferred or non-preferred (synonyms) compared with the selected term. These associated terms are also represented in the Thesaurus space by squares and their relationship to the selected term is represented by their relative proximity and opacity. Users can also use the checkboxes in the right-hand panel to show the terms in more than one language at once and to view scope notes for selected terms. When users decide to add a term to their query, they do so by clicking on its square in the centre of the screen, at which time it is added to the Summary of Terms list, or term pool, at the top of the right-hand panel. Users can add as many terms as they like, delete them at any time, choose to keep them in only one language rather than multiple languages, and combine them using the Boolean operators below the list. When they have finished formulating their query they click Retrieve Documents to view the results (Figure 3). The red dots in the middle around the green box represent the results retrieved for the chosen term. The Green box in the middle shows the thesaurus term and its French equivalent as well as the number of documents indexed using that term.

3 Methodology

Twenty-five participants from the University of Alberta were recruited for this study by purposive, maximum variation and snowball sampling. Although the participant pool included students and faculty members across departments, multilingual volunteers—particularly those from the Department of Modern Languages and Cultural Studies—were specifically targeted throughout the recruitment process. The resulting participant pool was diverse, comprised of professors, graduate, and undergraduate students from a variety of disciplines, including Applied Linguistics, Latin American Studies, French Language Studies, Romance Languages and Literatures, Library and Information Studies, Humanities Computing, English and Film Studies, Education, Chemical Engineering, History, Political Science, and Music. The group contained three professors, two doctoral students, seven master's students, and thirteen undergraduates. Twenty-three of these participants were women; two were men. Of these participants, thirteen were bilingual (seven spoke French fluently; two spoke Spanish; four participants respectively spoke Mandarin, German, Latin, and Russian). One participant spoke German (first language), English, French, and Latin. Six of these participants currently conduct research in more than

one language. This study used a wide range of data gathering tools, including pre-test, post-test and usability questionnaires; interviews; audio, video and screen capture; the think aloud technique and direct observation. Data from the interviews was collected verbally, digitally, and in written form. For the first 5-10 minutes of the interview, participants responded orally to a series of questions related to their academic background, the nature of their research, and their preferred online search tools. Participants' responses were recorded in written format by the interviewer. Next, the participants were given a brief overview of the usability study before being asked to complete three identical tasks on the Searchling interface and T-Saurus interface, respectively. The interface used first was alternated between the users, and users would move between interfaces as they completed first Task 1 (on either interface), then Task 2 (on either), and so forth. During this part of the session, which generally lasted for 25-45 minutes, participants were asked to verbally analyze Searchling and T-Saurus utilizing a thinkaloud protocol as they interacted with the interfaces and completed the required tasks. Furthermore, the users' physical interactions, dialogue, and mouse movements were recorded by the Silverback multimedia program (for video, sound, and screen capture), while the interviewer answered questions, provided hints if needed, and made written notes on the participants' engagement with, and comments on, the prototypes

4 Results

4.1 Tasks

All users chose a combination of browsing and searching strategies to carry out the three search tasks. Around half of the users chose to carry out a search first for their three tasks. The other half decided to use browsing strategies to find the term and its associated terms. However, browsing accounted for a significant part of their interaction, particularly for Task 1 in which users were asked to find the term 'Democracy' and one of its related terms. This task required that they interact with the thesaurus to browse and find a related term from among a list of terms that were hierarchically or semantically related to the term Democracy. In Searchling they typically decided to use the high level facets and the terms under each facet. Within T-Saurus, users browsed the alphabetical list on the left-hand side of the screen to find the term. A number of users liked Searchling for its results display as it showed the retrieved documents within the same interface without losing the context of thesaurus or search. In more than 10 searches, users found that the red-dot visualization representing the retrieved documents in T-Saurus was vague and at times difficult to interact with. In Task 2 users were asked to interact with multilingual and filtering (Boolean operators) features of both interfaces to combine two terms and retrieve doucments in Spanish. Due to the multi-term nature of this Task, almost half of the users conducted Boolean searches first and browsing next. The Boolean search features of both Searchling and T-Saurus were found very useful by participants. Several commented that that they would prefer an advanced search option built into the query formulation stage of the search process, where they could use a combination of Boolean operators. The auto-completion feature within the T-

Saurus search box was found particularly useful and interesting. A majority of users expressed positive comments about the search term pool feature available in Searchling. Also, when users browse and choose a term in the thesaurus, the selected term(s) gets automatically added to the search term pool area, making it particularly easy for the searcher to create a more sophisticated query statement. All users found the language option within Searchling flexible, intuitive and easy to use. The results from the third task were mixed. This was, in part, due to the wide variety of search terms that users employed to carry out searches based on their own specific research interests and needs. Some users experienced frustration as they were not able to find terms that matched their query terms. Others found specific features of each interface appealing or useful. In Task 3 users were asked to freely search or browse using their own information and research needs. Most comments made by users for Task 3 focused on various interface features, such as (in T-Saurus) the breadcrumb feature and visual grouping of thesaurus terms, and (in Searchling) the sort of results display, search term pool, and linear organization.

4.2 Multilingual Features in Searchling and T-Saurus

As was discussed before, both interfaces allow users to choose thesaurus terms for searching in three different languages, namely English, French and Spanish. Participants liked the language features in both user interfaces for their easy access and contextual display of thesaurus terms in different languages. The majority of study participants found the Searchling interface user-friendly, intuitive, and particularly flexible across the languages. Around 88% of users agreed or strongly agreed that the Searchling interface would help them locate relevant results in Spanish, French and/or English, whereas 72% of users agreed or strongly agreed that the T-Saurus interface provided useful language options. The main difference between the language feature in Searchling and T-Saurus is that the three languages in Searchling are all clickable and upon clicking on each thesaurus term, that term becomes prominent in bold and the equivalent terms in the other two languages will be shown. In T-Saurus the feature is different in that clicking on a thesaurus term will show the term in English with the other two equivalents. One user noted the generative attributes of the interface, stating that she would use Searchling "to find relevant information in English and French for a particular topic that she may have otherwise not thought of." Another user commented that showing English related terms when carrying out a French search would be very useful in Searchling. Above 60% of the users thought that using a thesaurus-enhanced search interface would help them formulate research questions.

4.3 Thesaurus and Search Functions

One of the research questions in this study was to examine how users evaluate the thesaurus and search functions of the two interfaces. We were interested to know what kind of thesaurus presentation and visualization would be easy to understand and easy to use by academic users. The Searchling user interface provides a design similar to a faceted search interface. It uses the high level facets of the UNESCO thesaurus along with a list of terms for each facet. The T-Saurus interface provides a more visualized and interactive interface where users have to interact with the interface to choose thesaural

term relationships, such as more general, more specific or related terms. Both Searchling and T-Saurus allow users to browse thesaurus terms both hierarchically and alphabetically. The default option for Searchling is the faceted view of the thesaurus, whereas in T-Saurus the alphabetical list is default. For assessing the affordance strength of the two user interfaces we asked the users to give the thesaurus and search functions of Searchling and T-Saurus a score ranging from Very Difficult to Very Easy, or Not at all to Very Much, depending on the question. The search and thesaurus functions in Searchling were rated higher (76% and 68% respectively) than T-Saurus (40% and 40% respectively), indicating that Searchling's search and thesaurus functions are significantly easier to use than those functions in T-Saurus. Also, 16% of users found that T-Saurus was more difficult to use than Searchling.

As part of the assessment of affordance strength, we asked users two additional questions. The first question was if the thesaurus-based grouping of results provided by these interfaces was helpful in developing searches. Around 24% of users mentioned "very much" while 52% said "somewhat". The second question asked them whether they would be motivated to use Searchling or T-Saurus as an interface to their more frequently used databases. It was found that around 72% would be Very much or Somewhat motivated to use Searchling, while only 52% would be motivated to use T-Saurus. A final question asked users which interface they generally preferred, Searchling or T-Saurus. Around 56% preferred Searchling, while 40% thought T-Saurus was a better user interface. This final questions confirms the findings related to all the interface features discussed before, namely multilingual and thesaurus and search functions.

5 Conclusion

This comparative usability study has yielded promising implications for the multilingual thesaurus-enhanced user interfaces to support users in their information seeking process. The visualization in both interfaces was found to be comprehensible to users. A common observation for both interfaces was that users found the thesaurus functions useful for broadening and narrowing down the scope of their research activities. In general, the Searchling interface was found to be more favorable and easier to use in terms of multilingual features, thesaurus and search functions and users' motivation to use such an interface for research purposes. Though T-Saurus, was preferred by fewer users than Searchling, the most promising finding for T-Saurus was that it has the potential not only to support browsing, searching, and query formulation, but also to transform these processes. It was found that linear thinkers preferred Searchling, whereas visual learners liked T-Saurus. Searchling is a linear, sequential and visual interface that uses faceted structure as its default interface and the thesaurus terms such as more general, more specific and related terms would be shown automatically as soon as a term was selected by the user. T-Saurus, on the other hand, provides users with a more interactive and dynamic visualization interface where users need to interact with and choose individual thesaurus term relationships to be shown. In general, this study found that faceted presentation of thesauri was more favourable than visual and graphical. The results from this study indicate that for exploring and using thesaurus terms in a search user interface, most users prefer related, more specific and more general terms to be shown along with the selected term without additional effort. In other words, upon searching for a term,

users should be provided all the related terms automatically for detailed view and selection. It would be interesting to observe how users with different cognitive, perceptual and learning styles may have different preferences when they interact with visual user interfaces. Further research may use a verbalizer/visualize cognitive inventory [10] to formally study how each learning style will affect users' interaction with visual user interfaces enhanced with such semantic tools as thesauri. Many users noted positive implications of the thesauri functions in both Searchling and T-Saurus, from undergraduates new to a topic to multilingual experts well versed in the terminology of their field of research.

Acknowledgement. The authors wish to acknowledge the Social Sciences and Humanities Research Council of Canada (SSHRC) for providing the funding support.

References

[1] White, R.W., Kules, B., Drucker, S.M., Schraefel, M.C.: Supporting Exploratory Search, Introduction, Special Issue. Communications of the ACM 49(4), 36–39 (2006)

[2] Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM 49(4), 41–46 (2006)

[3] Gordon, A., Domeshek, E.A.: Déjà vu: a knowledge-rich interface for retrieval in digital libraries. In: Marks, J. (ed.) Proceedings of the 1998 International Conference on Intelligent User Interfaces (IUI), San Francisco, January 6-8, pp. 127–134. ACM Press, New York (1998)

[4] Sutcliffe, A.G., Ennis, M., Hu, J.: Evaluating the effectiveness of visual user interfaces for information retrieval. International Journal of Human–Computer Studies 53(5), 741–763 (2000)

[5] McKay, D., Preeti, S., Hunt, R., Cunninghum, S.J.: Enhanced browsing in digital libraries: three new approaches to browsing in Greenstone. International Journal on Digital Libraries 4(4), 283–297 (2004)

[6] Blocks, D., Binding, C., Cunliffe, D., Tudhope, D.: Qualitative evaluation of thesaurus-based retrieval. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, pp. 346–361. Springer, Heidelberg (2002)

[7] Stafford, A., Shiri, A., Ruecker, S., Bouchard, M., Mehta, P., Anvik, K., Rossello, X.: Searchling: User-Centered Evaluation of a Visual Thesaurus-Enhanced Interface for Bilingual Digital Libraries. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 117–121. Springer, Heidelberg (2008)

[8] Yee, K., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 2003), Ft. Lauderdale, Florida, USA, pp. 401–408 (April 2003)

[9] Shiri, A., Ruecker, S., Fiorentino, C., Stafford, A., Bouchard, M., Bieber, M.: Designing a Semantically Rich Visual Interface for Cultural Digital libraries Using the UNESCO Multilingual Thesaurus. In: Sudweek, F., Hracovech, H., Ess, C. (eds.) Proceedings of Cultural Attitudes Towards Technology and Communication 2010 Conference: Cultural Diversity in e-Learning and/or m-Learning, June 15-18. University of British Columbia, Vancouver (2010)

[10] Leutner, D., Plass, J.L.: Measuring learning styles with questionnaires versus direct observation of preferential choice behaviour in authentic learning situations: the visualizer/verbalizer behavior observation scale (VV-BOS). Computers in Human Behavior 14, 543–557 (1998)

Multilingual Adaptive Search for Digital Libraries

M. Rami Ghorab[1], Johannes Leveling[2], Séamus Lawless[1], Alexander O'Connor[1], Dong Zhou[1], Gareth J.F. Jones[2], and Vincent Wade[1]

[1] CNGL, Knowledge and Data Engineering Group, School of Computer Science & Statistics, Trinity College Dublin, Dublin 2, Ireland
{ghorabm,seamus.lawless,alex.oconnor,dong.zhou, vincent.wade}@scss.tcd.ie
[2] CNGL, School of Computing, Dublin City University, Dublin 9, Ireland
{jleveling,gjones}@computing.dcu.ie

Abstract. We describe a framework for Adaptive Multilingual Information Retrieval (AMIR) which allows multilingual resource discovery and delivery using on-the-fly machine translation of documents and queries. Result documents are presented to the user in a contextualised manner. Challenges and affordances of both adaptive and multilingual IR, with a particular focus on digital libraries, are detailed. The framework components are motivated by a series of results from experiments on query logs and documents from The European Library. We conclude that factoring adaptivity and multilinguality aspects into the search process can enhance the user's experience with online digital libraries.

1 Introduction

Estimates show that 60% of World Wide Web (WWW) users are non-English speakers[1] and the content available in non-English languages is growing fast. The variety in users and language means that regardless of which languages a user speaks, there is a large volume of content which they cannot easily discover, consume or comprehend. Moreover, on the specific scope of European countries, information access systems that cater for European users must take into consideration the different linguistic and cultural backgrounds of their users. Online Digital Library (DL) portals, such as The European Library (TEL)[2] support a variety of languages by allowing the user to type in queries in different languages, select different interface languages, or view result documents written in different languages. However, in this paper we argue that there is yet more potential in exploiting adaptive techniques to improve the user's experience with multilingual search in DLs, for example through query adaptation and result adaptation.

In this paper, we describe a framework for Adaptive Multilingual Information Retrieval (AMIR) which comprises techniques from Information Retrieval (IR) and Adaptive Hypermedia (AH) [1]. Traditional IR approaches focus on the retrieval and presentation of documents in a ranked list, from which the user chooses the best documents to view. AH enables the aggregation of content in dynamically tailored hypertextual presentations, ensuring that results are adapted to the user's preferences, goals, and

[1] http://www.internetworldstats.com/
[2] http://www.theeuropeanlibrary.org/

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 244–251, 2011.

context. The AMIR framework supports multilingual resource discovery and delivery using on-the-fly machine translation of documents and queries. This allows us to bridge the gap between the user's language and document languages. Furthermore, adaptive techniques are applied to present the content that is most relevant to users. A key reason for combining AH and IR is that adaptivity can contextualise IR.

The framework contributes to three different strands of research: i) user-centric research, determining the user requirements and consequently the required functionality of the search system, content processing, and result presentation; ii) system-centric research, targeting the study and development of system components for indexing, annotation, retrieval and presentation of multilingual documents; and iii) evaluating the system and its components in user studies and formal evaluation benchmarks and investigating the combination of approaches from AH and IR.

The rest of this paper is organised as follows: Section 2 discusses related work; Section 3 describes the AMIR framework and presents component-level evaluations for result adaptation; and Section 4 presents conclusions and future work.

2 Related Work

A digital library (DL) constitutes a network of federated information sources, interfacing to different types of electronic documents made available by these sources. Users can submit queries to the DL to search for documents which are stored at the local repositories of one or more federated sources [10].

The number of DL research projects and systems indicates that there is growing interest in DL. For example, Europeana[3] is a virtual library providing access to millions of digital items, including music, films, and text. It aims to facilitate access to Europe's cultural and scientific resources. The MILE project[4] promotes European cultural heritage and makes images accessible by improving their metadata annotation. One of the major areas of investigation in MILE was concerned with improving search based on metadata. The MultiMatch search engine[5] focuses on information from cultural heritage institutions, identifying relevant documents regardless of the language. It can organise and display search results in an integrated, user-friendly manner, allowing users to access and exploit the retrieved information regardless of language barriers. The CACAO Project[6] offers an integrated approach for accessing, understanding and navigating multilingual documents in library catalogues, enabling users to better exploit electronic content. Three thematic portals were created as part of the project, realising multilingual book search on History, Mathematics and Geography. ezDL[7] (formerly DAFFODIL) is a user-oriented front-end for DLs which supports proven search strategies, integrating different DLs [5]. DAFFODIL was adaptive towards different user wishes, regarding preferences concerning content and system involvement. This was achieved via a profile of user's interests, created from items in a personal library [4].

[3] http://www.europeana.eu/portal/
[4] http://www.mileproject.eu/
[5] http://www.multimatch.eu/aboutmultimatch.tml
[6] http://www.cacaoproject.eu/
[7] http://ezdl.de/

TEL offers access to digital and bibliographical resources of major European national libraries. TEL provides a virtual collection of information resources from many domains. The LADS task (Log Analysis for Digital Societies), which is part of the Log-CLEF track at CLEF[8] aims at investigating user actions in multilingual search systems. The LADS task is involved with different experimental datasets, including log files of user interactions with TEL.

While the AMIR framework presented in this paper can be customised for search in DLs, it is not restricted to a particular search domain; the framework can in fact be configured to run on different corpora, and it also allows plugging in alternative components to carry out translation, query adaptation, and result adaptation. The following sections discuss the components of the AMIR framework. Moreover, for two of the components, a component-level evaluation is presented.

3 Framework for Adaptive Multilingual IR

The motivation for the AMIR framework originates from the observations that: i) users from different linguistic or cultural backgrounds behave differently in search; ii) there are identifiable patterns in user actions; and iii) user queries and action patterns reflect the mental model or prior knowledge of a user about a search system [3]. These observations indicate a need to address adaptivity and multilinguality aspects in DL search.

The AMIR framework architecture is shown in Figure 1. The framework is fully functional. It is implemented in Java and follows the Model-View-Controller (MVC) architecture. The framework contains components of the following types: 1) controllers which govern the sequence of execution of the retrieval and adaptation processes; 2) components which hold the algorithms for query adaptation, result list merging, and result list adaptation; and 3) components which deal with external services such as web search engines, machine translation web services and automatic language identification. The components of the framework support a spectrum of functionality regarding multilingual and adaptive IR. The framework components are illustrated below.

The *User Modelling* component gathers user and usage information about the users and organises this information in user models. This component can make use of both, implicit and explicit information gathering approaches in order to populate the models. The implicit approach involves processing the user's search history. The explicit approach involves web forms that ask users to engage in two activities: 1) supplying personal/demographic information; and 2) scrutinising the inferred search interests in the user model. The information from the user model is used to adapt the queries, the results, the interface language, and the appearance of the search application.

In the *Query Adaptation & Translation* component, the user's query is adapted in two stages: pre-translation query expansion and post-translation expansion. Query expansion can be based on terms obtained from blind relevance feedback, the user model, or both [2]. Blind relevance feedback involves extracting terms from top-ranked documents retrieved with the original query or with the translated query. These terms are then added to the query to encourage retrieval of additional relevant results.

[8] Cross-Language Evaluation Forum: http://www.clef-campaign.org/

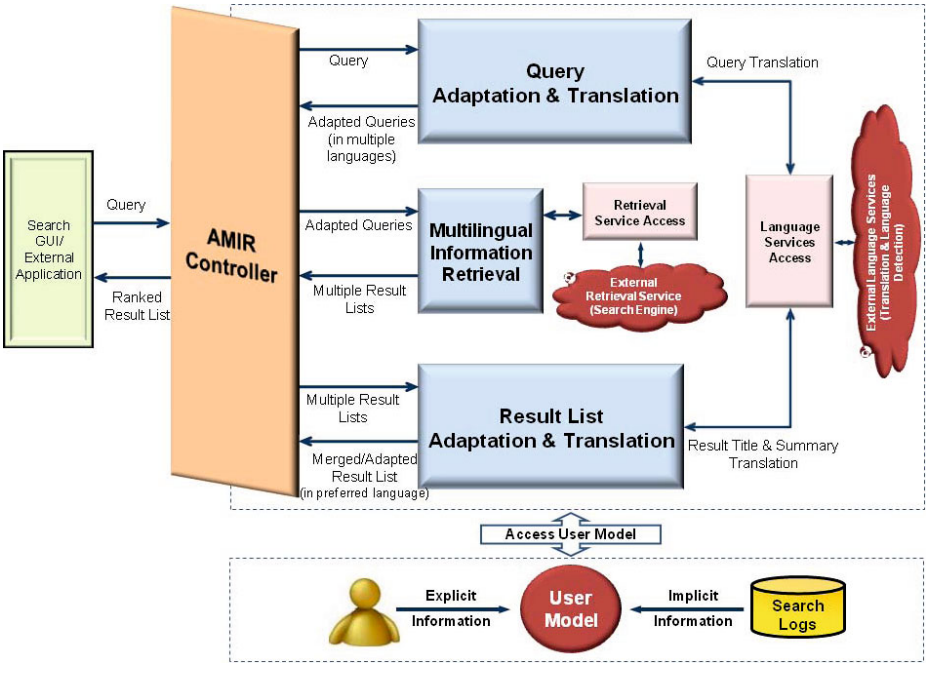


Fig. 1. Framework Architecture

The *Multilingual Retrieval* component performs multiple monolingual retrieval runs, one for each target language, and then passes multiple result lists to the Result List Adaptation & Translation component. Currently, the framework supports automatic translation of queries and documents using web services such as Google Translate[9] or MicrosoftTranslator[10]. In addition, translation by OpenMaTrEx[11], an open-source machine translation toolkit, is supported for domain-specific translations.

The *Result List Adaptation & Translation* component merges results from multiple languages into a single list which is then translated to the user's preferred/native language. Furthermore, this component performs result list re-ranking, which can be based on latent semantic information induced from the documents or based on the user's search interests. In addition, in environments where the user is presented with a list of candidate collections that group the results together (such as in TEL, where users can select library collections in the results page), the component can re-rank the list of collections based on the user's language or country. Given the importance of this component with respect to DLs, result adaptation approaches are discussed in more detail in the next subsections.

[9] http://translate.google.com/
[10] http://www.microsofttranslator.com/
[11] http://openmatrex.org/

3.1 Collection Re-Ranking Based on User Models

In several DL web sites, the user is presented with search results that are grouped under different target library collections. Our collection re-ranking evaluation aimed at improving retrieval effectiveness by ordering the list of collections based on user information, so that collections that match the user's language or country are ranked higher in the list. We evaluated the collection re-ranking component as part of our participation in the LogCLEF track of CLEF 2010 [7]. The evaluation experiments were based on search logs of TEL where all user interactions with the portal (including user queries and clicks on collections) were recorded [8].

The user interface of TEL shows a list of collections on the left side, which is by default sorted in alphabetical order. A list of results from the selected collection is shown on the right side, In TEL, a collection is either a library catalogue or an online resource. For the collection re-ranking experiments, we associate TEL collections with languages, based on the official languages that are spoken in the collection's country. A subset of the TEL interaction logs was extracted, which contains submitted queries and clicked collections during the month of February 2007. This included approximately 566 queries from different languages.

In order to evaluate the retrieval precision over the list of collections presented to the user, we used the collections that the user clicked on as implicit relevance judgements (i.e. binary relevance assessments where the clicked collections are assumed to be the relevant ones). However, the notion of relevance in our experiments is in terms of matching language and country. Mean Average Precision (MAP) was used for evaluation where the MAP score was calculated across the queries in the selected subset of the data. The original ranked list of collections (i.e. the one presented to the user by TEL) was used as the baseline for evaluation of retrieval precision (0.580 MAP). Several alternative re-ranked lists were investigated and compared to the baseline.

The collection re-ranking builds on three attributes: country (location from which the query was submitted), query language, and interface language. Collection scores are computed as a weighted linear combination of attributes with the collection's country and language as follows ($M_X = 1$ if two items match, $M_X = 0$, otherwise): i) M_c: *user's country* and collection's country match; ii) M_q: *query's language* and collection's language match (i.e. a match with any of the official languages spoken in the corresponding country); iii) M_i: *interface language* and collection's language match.

Each of the above attributes is multiplied by a scalar weight (W_c, W_q, W_i respectively) to control the degree of contribution of each attribute in the function. The collection list is re-ranked based on descending order of the new collection score. Table 1 shows the MAP values for some selected re-ranking runs with different weights that ranged from 0.0 to 1.0. The results showed a significant improvement of 27.4% MAP for the re-ranked collection lists (with weights: $W_c = 0.1$, $W_q = 0.3$, $W_i = 0.6$) over the baseline ranking. This improvement is statistically significant as per the t-test (with p=0.01). The results of this experiment suggest that there is opportunity for improving the user's experience with multilingual search in DLs if the user's language and country are taken into consideration when adapting the results.

3.2 Multilingual Result Re-Ranking Based on Query-Document Features

Multilingual result adaptation can be achieved by re-ranking an initial result set. The simplest approach is to directly apply the monolingual methods on the results obtained using a translated query. The drawback of this approach is that translation errors will be propagated to the adaptation process, which may result in unsatisfactory performance. In previous work we proposed a multilingual re-ranker component which incorporates scores generated using external knowledge to enhance the semantic space produced by the latent concept method [11] through a linear combination model. This method is designed to solve the multilingual document re-ranking problem by automatically inducing a semantic correspondence between two languages (query language and document language) using parallel corpora as training data. For the experiments described in this paper, a parallel corpus extracted from Wikipedia data was employed. The correspondence between two languages was then used to project the query into another language in the semantic space to accomplish the re-ranking task. Formally, the latent semantic space was produced by a Latent Dirichlet Allocation (LDA) model.

The TEL corpora from CLEF 2009, which contain documents in English, French and German, were employed for evaluation of this component. These collections were chosen to test the scalability of the proposed method in different settings and over different languages. The corpora were provided together with queries and relevance judgments by the organisers of the CLEF ad-hoc retrieval task. Wikipedia documents in English, French and German were used as an explicit concept space. Only those articles that are connected via cross-language links between all three Wikipedia databases were selected. The results suggest that our method outperforms a baseline ranking system (standard BM25) and a previous proposed re-ranking method that only uses latent features by a statistically significant margin in many test runs. The experiments also confirm that directly applying monolingual methods into the cross-lingual applications may not always produce the most beneficial results. The results for two language pairs (French-English and German-English) are shown in Table 2 ($*$ indicates significant improvement over the initial retrieval). The following IR evaluation metrics were used to evaluate retrieval precision: precision at N documents (P@N), Normalised Discounted Cumulative Gain (NDCG), and MAP.

Table 1. MAP for re-ranking with different weight combinations. MAP for the baseline is 0.58

	Weights (W_c, W_q, W_i)						
W_c	0.1	0.1	0.6	0.6	0.0	0.0	1.0
W_q	0.3	0.6	0.1	0.3	0.0	1.0	0.0
W_i	0.6	0.3	0.3	0.1	1.0	0.0	0.0
MAP	**0.74**	0.73	0.72	0.71	0.71	0.68	0.60
Improvement [%]	**+27.4**	+25.3	+23.9	+22.2	+22.2	+18.1	+3.4

Table 2. Multilingual Re-ranker

	French-English			German-English		
metric	baseline	latent	multilingual re-ranker	baseline	latent	multilingual re-ranker
P@10	0.444	0.470*	0.474*	0.416	0.466*	0.470*
P@20	0.389	0.389	0.391	0.364	0.389*	0.394*
NDCG	0.368	0.370	0.371	0.362	0.372*	0.372*
MAP	0.212	0.216	0.218*	0.210	0.223*	0.224*

3.3 The Evaluation Gap: User Studies vs. IR Evaluation

Contextual information about the user, content and environment is increasingly being used to support the tailored delivery of information in IR. The personalised discovery, retrieval and presentation of content can provide an enhanced information seeking experience to the user. While such tailored experiences can produce a more informative response than a traditional ranked list approach, there are many challenges associated with evaluating these approaches. As AMIR and response composition become more widely used, traditional approaches to IR evaluation may become less effective or applicable in isolation. The complex functionality offered by these systems and the variety of users who interact with them, mean both component-level evaluation and extensive user-based evaluation are required to comprehensively assess the system's performance. We have proposed an evaluation approach which combines and enhances evaluation methodologies from both the AH and IR communities [6]. In order to sufficiently evaluate both the adaptive functionality and the retrieval effectiveness of an AMIR system, a hybrid approach is necessary. This involves user-centric assessment, layered evaluation of the adaptivity which has been applied, and quantitative performance metrics relating to the content delivered.

IR has traditionally been evaluated using precision and recall and derivatives of these metrics such as MAP and the F-measure. These metrics measure the accuracy and scope of the retrieval of relevant documents. While these metrics are valuable in measuring the effectiveness of real world tasks, they are more typically used to evaluate retrieval effectiveness with test collections in laboratory IR experimental settings.

Numerous measures for the evaluation of adaptivity in adaptive systems have been proposed [9] which aim to measure the scientific performance of components and perform user-based evaluation of the adaptivity offered by the system. The approaches can be broadly divided into three categories: i) adaptivity metrics; ii) user-interaction metrics; and iii) performance metrics. These form a set of necessary elements of a hybrid AMIR evaluation model which can be defined irrespective of the system being evaluated. The key challenge is to be able to adequately combine the data-driven approach to IR evaluation with the more user-focused approach to evaluation from AH.

4 Conclusions and Future Work

In this paper we proposed a framework that caters for adaptivity and multilinguality aspects in DL search. We have presented component-level evaluations and highlighted

evaluation aspects with respect to the fields of IR and AH. We showed that there is scope for improving the effectiveness of search in DLs if user and usage information is exploited with respect to the multilingual dimension. While integrating adaptivity and personalisation capabilities into search systems is desirable, the evaluation approach will have to be revised by combining evaluation approaches from AH and IR, where evaluation focuses on user and usability aspects in conjunction with retrieval precision.

The current version of the AMIR framework focuses on adaptivity in a multilingual context by adapting the search system to a group of users. As part of the future work, the AMIR framework will aim at personalisation of information access by integrating different recommender systems for individual users, thus tailoring the results for a query to single users.

Acknowledgments. This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (http://www.cngl.ie/).

References

1. Brusilovsky, P.: Adaptive hypermedia. User Modeling and User-Adapted Interaction 11(1), 87–110 (2001)
2. Chirita, P.A., Firan, C.S., Nejdl, W.: Personalized query expansion for the web. In: SIGIR 2007, pp. 7–14. ACM, New York (2007)
3. Ghorab, M.R., Leveling, J., Zhou, D., Jones, G.J.F., Wade, V.: Identifying common user behaviour in multilingual search logs. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 518–525. Springer, Heidelberg (2010)
4. Gövert, N., Fuhr, N., Klas, C.P.: Daffodil: Distributed agents for user-friendly access of digital libraries. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 352–355. Springer, Heidelberg (2000)
5. Kriewel, S., Fuhr, N.: Adaptive search suggestions for digital libraries. In: Goh, D.H.L., Cao, T.H., Sølvberg, I., Rasmussen, E.M. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 220–229. Springer, Heidelberg (2007)
6. Lawless, S., O'Connor, A., Mulwa, C.: A proposal for the evaluation of adaptive personalised information retrieval. In: CIRSE 2010, Milton Keynes, UK (2010)
7. Leveling, J., Ghorab, R., Magdy, W., Jones, G.J.F., Wade, V.: DCU-TCD@LogCLEF 2010: Re-ranking document collections and query performance estimation. In: CLEF 2010 LABs and Workshops, Notebook Papers, Padua, Italy, September 22-23 (2010)
8. Mandl, T., Agosti, M., Di Nunzio, G.M., Yeh, A., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: The CLEF 2009 multilingual logfile analysis track overview. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 508–517. Springer, Heidelberg (2010)
9. Raibulet, C., Masciadri, L.: Evaluation of dynamic adaptivity through metrics: An achievable target? In: WICSA/ECSA 2009, pp. 341–344 (2009)
10. Rigaux, P., Spyratos, N.: Metadata inference for document retrieval in a distributed repository. In: Maher, M.J. (ed.) ASIAN 2004. LNCS, vol. 3321, pp. 418–436. Springer, Heidelberg (2004)
11. Zhou, D., Lawless, S., Min, J., Wade, V.: A late fusion approach to cross-lingual document re-ranking. In: CIKM 2010, pp. 1433–1436. ACM, New York (2010)

Making Sense in the Margins: A Field Study of Annotation

James Blustein[1,2], David Rowe[2,3], and Ann-Barbara Graff[2,4]

[1] Dalhousie U. (Faculty of Computer Sci. & School of Info. Mgmt.)
jamie@ACM.org
[2] Hypertext Augmenting Intelligent Knowledge Use (HAIKU) Project
[3] Dalhousie University (Faculty of Computer Science)
[4] Nipissing University (English Studies)

Abstract. We report on three years of data collected in the field from students in graduate and undergraduate seminars at two universities. The students annotated texts for discussion in classes where hypertext and computer interfaces were core topics. The results of our analysis show how annotation style changes with a combination of experience and study of material related to annotation. Our major conclusions are that there are essentially six purposes for scholarly user-readers to annotate; and support for textual glosses is a necessary part of any successful annotation technology for such use. Our study suggests tools that will be appreciated by e-text users.

1 Introduction

In academia, annotation is a means of making sense of complex material, marking engagement, navigating, and establishing a foothold for original thought. Wolfe and Neuwirth [27, p. 338] note that 'empirical research involving students suggests that annotations improve comprehension, facilitate rereading and reviewing of documents, and help writers bridge reading and writing practices'. To date, computer-based annotation tools lack something, insofar as they have been found to be less effective than their paper analogues for some purposes [20,15]. Athough screen technologies are becoming more like paper other limitations, in particular the user interfaces of annotation software, impair wide-spread use [8].

We present an outline of the techniques that scholarly reader-users currently employ. Although annotation has been studied historically (e.g. by Jackson [12] and Hauptman [10]), practices have changed in recent generations. Mangen [15, p. 404] cites recent articles suggesting that 'reading modes and habits in general are changing due to steadily increasing exposure to digital texts'. Annotation however does not seem to have been affected by such exposure [20,25].

2 Method

Rather than relying on prescriptive or historically derived taxonomies to identify engagement we follow Marshall and Brush's [20] lead in analyzing actual annotations created for known purposes and with recognizable motives.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 252–259, 2011.

We gave graduate and undergraduate students printed readings on topics related to hypertext with the instruction to prepare to discuss them in seminar in the following week. The only difference between these readings and regular ones was that these were presented in specially prepared form. Students were encouraged, but not required, to make annotations on the documents.

Materials. All of the readings were presented on large pages[1] with wide margins, and ancillary material (e.g. diagrams and definitions) beside the text proper.

One reading was a collection of documents arranged around two related selections from a popular monograph about isochrestic design [22,23]. The other reading was about the use of hypertext in education [21].

Participants and Activities. The graduate students were enrolled in CompSci, Library and Information Studies, or Electronic Commerce at Dalhousie Univ. For many, English was not their first language. The undergraduates were enrolled in Literary Studies at Nipissing Univ. Of the 55 students enrolled between 2006 and 2009 fifteen volunteered to participate in the study ($N = 15$).

Students were graded on their discussion of the readings. Grad students were also graded on written critical assessments of the readings. Participation was not anonymous but the prof could not learn who had done so until after the course.

The undergrads studied Lipking's essay on marginal gloss [14] and Coleridge's annotated *Rime of the Ancient Mariner* [5] before the exercise. The only document they read for this study was the article about hypertext in education [21].

The grad students studied articles about annotation over the next weeks before a second document was circulated. They received grades and comments for their summaries of the first document before they could annotate the second.

Coding Method. The annotations were categorized according to a taxonomy based on studies by Marshall et al. [18,17,20]. Each individual annotation was examined and its features (see below) were recorded. For annotations that were hybrid types each component was counted separately, i.e., if an annotation was composed of an arrow and a box then 3 annotations were recorded: compound, arrow, and box. The same procedure was used when tallying annotation function.

Annotation Category. Annotations were coded as either telegraphic or explicit. A *telegraphic annotation* is a non-text based marking. Underlining, highlighting, and asterisks are examples of telegraphic annotations. An *explicit annotation* is a textual note (from single words to paragraphs).

Audience. All annotations in our study were *private*, i.e. not intended to be read by others. Annotations for a *public* audience tend to be much lengthier and less spontaneous — they are not about working towards understanding but a performance to demonstrate mastery to later readers [19].

Location. The place on the document where the annotation was recorded is its location. Annotations that do not overlap the text proper or otherwise obscure it are *external*. Such annotations may be in the margins or on separate pages.

[1] Ledger-size paper is 11"×17" — double the width of the letter-size ($8\frac{1}{2}$"×11") paper with which the students in North America were used to working.

Annotations such as highlighting and circles around words (which lie over the text proper) or writing which is interlinear we call *within-text*.

Type. The *type* of an annotation is a characterization of the physical mark the user-reader leaves on the document. E.g.: textual notes, shapes drawn around the text proper, underlining, asterisks, arrows and other deictic devices.

Following Bradley and Vetch [3, p. 226] we **further** classified marks as casual or meaningful. *Casual marks* were incidental, underlining or highlighting. All others were considered *meaningful*. This distinction follows observations by Charney [4] and others who have found that successful reader-users of hypertexts are so-called *active readers* in Adler and van Doren's [1] sense: they consciously make meaning from text. *Passive readers* read superficially with little or no cognition. Passive readers do not analyse the text.

Readers of both extreme types (passive and active) announce their presence in texts by making marks. Those marks are personal, subjective, and temporal. They are reflective of the readers' most active yet potentially fleeting thoughts. Passive readers' marks are superficial or only used to mark their progress through the text. Active readers sometimes make such marks (i.e., when interrupted during reading or pondering a passage) but most active readers' marks are richer in information.

Type Category. The type category is only about the signs user-readers make. We classify the marks users make by type (anchor and content) and function. Function follows in a later category.

Jackson [12, p. 81] noted the 'essential and defining character of the marginal note ... is that it is a responsive kind of writing permanently anchored to preexisting written words'. *Anchor* type annotations serve to call attention or ascribe significance to the part of the document where they are located [20]. Highlighting and underlining are examples of annotation as anchor. *Content* type annotations are notes (drawings, text, etc.) which help reader-users concentrate on parts of the text they find important.

If anchors help the user-reader keep their bearings (in what Dillon [7] calls the information space of the document) then content type annotations orient the reader-users in the argument or draw attention to what they find key. Individual readers have ways of ranking their own marks in part because they generally use a limited repertoire.

Annotations that combine anchor and content types are *compound*s.

Function. We classify annotations by their ostensible purpose. Jackson [12, pp. 90,82] observes that not all readers are annotators; the annotator 'acts on the impulse to stop reading for long enough to record a comment'. We categorize the purpose of annotations into 6 classes by the level of engagement following Bloom's taxonomy[2] and Jackson's observations. This order

[2] Bloom's (recently revised) taxonomy is a standard ranking of the levels of learning [11]. The top four of the six levels (namely creating, evaluating, analysing, and applying), suggest the value of the use of natural language. The bottom two (understanding and remembering) do not require the use of language.

reflects the urgency, and possibly the complexity, of what readers must do to grasp the text:

1. *Interpretive* marks are made when users truly make the text their own: they add some of their own thoughts. An example could be a short note.
2. *Problem-working* marks most often appear near charts or equations, and suggest or record the reader's attempt to understand what is represented or expressed. Definitions are examples.
3. *Tracing progress* may be signaled by the highlighting of lengthy passages, indicating the reader may be overwhelmed by the text, or unable to recognize the relative importance of passages.
4. *Procedural* annotations are intended to draw the user-reader back to parts of the text that require further attention. An asterisk marking a particular sentence could be classified as procedural, for example.
5. *Place-marking and aiding memory* annotations indicate places where the reader signals their presence but not what they are thinking. Highlighting or circling of keywords are examples.
6. *Incidental markings* (e.g. doodles) seem to mark a lack of engagement.

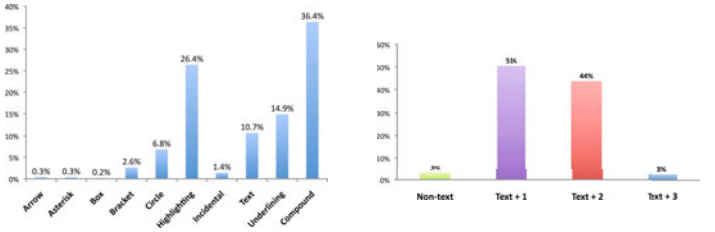
The examples, of course, are the general case as user-readers have their own set of idiosyncratic marks. In our study we found that all of the specific marks correspond to the six categories above.

Statistical Method. We use a mixed model repeated measures design. Of the fifteen participants, only five volunteered their annotations of both documents. For most we use within-subject ANOVAs which compare multiple measures of each participant's data. Because there were so few participants in both sessions, we computed a between-subjects analysis of 2 groups (one per session). The first *group* was students who participated in only the first session.

3 Results

Figure 1 shows that most of the annotations were compounds and that almost all of those include some textual annotation. Table 1 shows the distributions of annotations by count, type category and function. Table 2 shows the distribution of subtypes of marks across all sessions and users; Tab. 2(b) shows the number made by participants who completed both sessions.

Annotation is Idiosyncratic — by Count and Use. We found no difference between number of annotations made between sessions by the students who participated in both ($t = 0.347, df = 4$). However there were differences between the number of annotations used by the participants taken as a whole ($F(1, 14) = 44.10, p < 0.001, \eta_p^2 = 0.759$). There was a difference between the number of uses of the types of marks ($F(8, 112) = 7.72, p < 0.001, \eta_p^2 = 0.355$) as shown in Fig. 1a. The use of annotations differs by category ($F(2, 28) = 22.599, p < 0.001, \eta_p^2 = 0.759$) and function ($F(2, 28) = 19.741, p < 0.001, \eta_p^2 = 0.585$).



(a) All Types of Annotation (b) Components of Compounds

Fig. 1. Distribution of Types of Annotation

Table 1. Types of Annotations Used ($N = 15$)

(a) Marks Used

Mark	Mean	S.D.	Mark	Mean	S.D.	Mark	Mean	S.D.
Arrow	7.73	9.00	Asterisk	5.80	10.34	Box	0.07	0.26
Bracket	5.20	6.86	Circle	4.73	6.68	Highlighting	20.27	26.18
Incidental	0.53	0.92	Text	25.47	15.81	Underlining	10.67	13.75

(b) By Category

Category	Mean	S.D.
Compound	51.53	33.86
Explicit	5.73	6.05
Telegraphic	23.20	16.13

(c) By Function

Function	Mean	S.D.
Anchor	22.53	14.32
Compound	52.87	37.26
Content	5.07	5.44

Table 2. Comparison of Class of Annotations by Session

(a) *Between*-Subjects

	All ($N = 15$)		**First Session** ($N = 7$)		**Second Session** ($N = 8$)	
	Casual	**Meaningful**	**Casual**	**Meaningful**	**Casual**	**Meaningful**
Mean	31.47	49.00	23.00	44.00	30.25	36.63
S.D.	23.18	27.77	16.26	29.70	25.27	21.00

(b) *Within*-Subjects ($N = 5$)

	First Session			**Second Session**		
	All	**Casual**	**Meaningful**	**All**	**Casual**	**Meaningful**
Mean	24.4	13.80	26.80	22.4	20.20	31.80
S.D.	12.4	8.58	15.64	14.6	19.31	20.14

Casual and Meaningful Marks. A two-tailed paired within-Ss t-test shows a difference between meaningful and casual marks. More marks were casual than meaningful ($t = -3.335, df = 14, p < 0.005$). Within-Ss analysis showed the number of meaningful versus casual marks differed by participant but not group ($F(1, 13) = 12.568, p = 0.004, \eta_p^2 = 0.492$).

Table 3. p-values of Pairwise Differences in Types of Marks Across Sessions

	Incidental	Problem working	Procedural
Interpretive	< 0.001	0.001	0.001
Place marking	0.001	0.002	0.001

Difference over Time. A within-Ss analysis of session and function found no effect of function but a strong session effect ($F(4, 56) = 26.099, p < 0.001, \eta_p^2 = 0.651$). The strong demarcation between functions is evident from Tab. 3.

4 Discussion

Reading is a complex activity [6, p. 4] and yet we found that annotation by scholars for study fits into six basic modes (to use Mangen's [15, p. 404] term). Annotation is clearly idiosyncratic. Marks are personal, subjective, and likely temporal. The use of marks which reveal engagement differs between readers too. Some readers use mostly casual marks and others use many more thoughtful markings. Almost all at some time use what we term 'meaningful' annotation styles that show their engagement with the text.

Variety of Marks for a Few Purposes For every user-reader in our study, the variety of marks was small. As few kinds of marks are used their function should be easily supported by a limited palette of types of mark. Perhaps variation in meaning or significance could be represented by variation in colour, shape, etc.

Importance of Textual Annotation. Figure 1 and a statistical analysis not presented here strongly indicate that textual marking is a primary form of annotation. There are other strategies for marking presence but serious engagement can only be through words for '[w]hat a [person] cannot state he does not perfectly know'[3]. Given this it is curious that e-book systems still make it difficult for reader-users to make such notes. Perhaps readers' engagement with e-text would increase if writing — as distinct from keyboard use — were adopted[4].

5 Conclusion and Future Directions

As digital culture eclipses print culture, and as hypertext becomes the dominant medium of publication, the kinds of questions to be asked about annotation and marginal glossing are changing. Documents are not merely available on-line (that is to misunderstand the paradigmatic shift); documents on-line reflect a

[3] Quoted (by Gowers [9]) from *The Report of the Departmental Committee on the Teaching of English in England.* H.M. Stationery Office, 1921.

[4] Some have claimed that annotation mediated by keyboards is inferior to annotation with styli because of the parts of the brain that are involved [16], while others conclude that all forms of note-taking require substantial cognitive effort [24].

reconceptualising of text. Notions of permanence attached to the written word are thought of as fetish; palimpsests (literally the residuum of erased text on parchment, metaphorically textual edits thought of as obscured in a final draft) are now marked by digital traces and tags. Accordingly the ways that readers can mark their unique engagement and strategies of annotation are changing. However we must be mindful of what they do currently so that we can support the reasons, if not all of the 'intuitive' or familiar forms [13].

Classification of Annotation Styles of Scholarly User-Readers. The classification of marks we developed in our study (§2) accounts for every mark found in the study. A future direction will be to validate or correct that classification by applying it to a wider range of contemporary annotated documents. Retrospective self-analysis or talk-aloud studies are necessary to corroborate our assessment of user-readers purposes in making such marks.

Telegraphic Marks. Since the uses of telegraphic (i.e., non-textual) marks are quite limited (although they are certainly idiosyncratic) e-reader tools need only provide a small palette of such marks in several variations.

Textual Marks. Annotation captures a person struggling to make meaning and sense. When the user-reader is confronting a new idea, synthesizing it, or capitalizing on it then the engagement must be textual, i.e. with words.

It is clear that support for textual glosses is a necessity for the success of any annotation system for scholars. Precisely how textual annotation should be supported is unclear. Wolfe [26] and Black et al. [2] have shown the importance of simultaneous on-screen presentation of notes that do not obscure the original text. It is not yet clear which potential methods are best. A major distinction in current methods is whether glosses are present when readers view the text proper or if users must act to display the gloss [3, 27].

It is clear that annotation is sufficiently key to the experience of reading that interfaces must be designed to ensure that readers can continue to annotate texts. Knowing why people make annotative marks is more important than knowing precisely which marks they make. Of particular importance are textual marks as distinguished from figural marks. To support users' needs digital systems must support the functionality people seek from traditional tools but not necessarily ape users' methods.

Acknowledgments. We thank Patricia Oprea for research assistance. We are grateful to Dana Murphy for guidance and assistance with statistical analyses. SSHRC funded this research through an ASU grant administered by Nipissing.

References

1. Adler, M.J., van Doren, C.: How to Read a Book: The Classic Guide to Intelligent Reading. Simon & Schuster, Inc., Toronto (1972)
2. Black, A., Wright, P., Black, D., Norman, K.: Consulting on-line dictionary information while reading. Hypermedia 4(3) (1992)

3. Bradley, J., Vetch, P.: Supporting annotation as a scholarly tool—experiences from the online Chopin Variorum edition. Lit. and Ling. Comput. 22(2), 225–241 (2007)
4. Charney, D.: The effect of hypertext on processes of reading and writing. In: Selfe, C.L., Hilligoss, S. (eds.) Literacy and Computers, The MLA (1994)
5. Coleridge, S.T.: Rime of the Ancient Mariner. Dover, New York (1970); [The 1817 edition (with marginal gloss); illustrated by Gustave Dore (1870)]
6. Crowder, R.G., Wagner, R.K.: The Psychology of Reading: An Introduction, 2nd edn. Oxford U. Press, Oxford (1992)
7. Dillon, A.: Spatial-semantics: How users derive shape from information space. JASIS 51(6), 521–528 (2000)
8. Golovchinsky, G.: Reading in the office. In: BooksOnline 2008, pp. 21–24 (2008)
9. Gowers, E., Greenbaum, S., Whitcut, J.: The Complete Plain Words. Penguin (1987)
10. Hauptman, R.: Documentation. McFarland & Co., Inc. (2008)
11. Hoffmann, R., McGuire, S.: Learning and teaching strategies. Amer. Scientist 98, 378–382 (2010)
12. Jackson, H.J.: Marginalia: Reader's Writing in Books. Yale U. Press (2001)
13. Kelly-Bootle, S.: It takes two to intuit. Comp. Lang. (August 1989)
14. Lipking, L.: The marginal gloss. Critical Inquiry 3(4), 609–655 (1977)
15. Mangen, A.: Hypertext fiction reading: haptics and immersion. J. Res. in Reading 31, 404–419 (2008)
16. Mangen, A., Velay, J.L.: Digitizing literacy: Reflections on the haptics of writing. In: Zadeh, M.H. (ed.) Advances in Haptics ch. 20, pp. 385–401. InTech (2010)
17. Marshall, C.C.: Annotation: from paper books to the digital library. In: Witten, et al. (eds.) ACM Dig. Libs., pp. 131–140 (1998)
18. Marshall, C.C.: Toward an ecology of hypertext annotation. In: Grønbæk, et al. (eds.) ACM HT, pp. 40–49 (1998)
19. Marshall, C.C.: Reading and writing the electronic book. Morgan & Claypool (2009)
20. Marshall, C.C., Brush, A.J.B.: Exploring the relationship between personal and public annotations. In: JCDL 2004 (2004)
21. McKendree, J., Reader, W., Hammon, N.: The "homeopathic fallacy" in learning from hypertext. Interactions ii(3) (July 1995)
22. Molotch, H.: Leaps and visions. In: Where Stuff Comes From, pp. 68–71. Taylor & Francis, Abington (2003)
23. Molotch, H.: The semiotic handle. In: Where Stuff Comes From, pp. 25, 82–84
24. Piolat, A., Olive, T., Kellogg, R.T.: Cognitive effort during note taking. App. Cog. Psych. 19, 291–312 (2005)
25. Qayyum, A.: Analysing markings made on e-documents. Can. J. Inf. & Lib. Sci. 32(1/2), 35–53 (2008)
26. Wolfe, J.: Annotations and the collaborative digital library. Int. J. of CSCL 3(2), 141–164 (2008)
27. Wolfe, J.L., Neuwirth, C.M.: From the margins to the center: The future of annotation. J. Bus. and Tech. Comm. 15, 333–371 (2001)

One of These Things Is Not Like the Others: How Users Search Different Information Resources

Dana McKay[1] and George Buchanan[2]

[1] Library, Institute for Social Research, Swinburne University of Technology
P.O. Box 218 John Street, Hawthorn, VIC 3122, Australia
dmckay@swin.edu.au
[2] Centre for HCI Design, City University
Northampton Square, London EC1V 0HB, UK
george.buchanan.1@city.ac.uk

Abstract. Transaction log analyses are common practice to understand user behavior in both online databases and library catalogues. While there has been significant work done in each of these domains, there is little work comparing user queries between library catalogues and online resources. In this paper we report on an exploratory comparison between searches performed via the same interface in three different search systems: a library catalogue, an online research database, and Google Scholar.

Keywords: User behavior, search behavior, search interfaces, libraries.

1 Introduction

Users' behaviors when engaged in interactive search have been the subject of extensive study. Previous research has made clear that information seekers choose different resources for meeting different information needs; depending on the nature of their need, information seekers may look to the library for books or articles, they may search the web, or they may just ask a friend [1, 2]. Some of the selection process is based on where in the information seeking process users are: according to McKenzie, users have four stages of information seeking, only one of which is likely to find users actively engaging with information resources [3]. Similarly Kuhlthau and Marchionini defined search only as a small part of the information seeking process [4]. Ill-defined or amorphous information needs are refined by browsing [5, 6], encountering information [3] or, in a library setting, by using a reference librarian as an intermediary [7]. With increasing availability of online information, though, the opportunity to browse is less and less common, and search is becoming the dominant information-seeking paradigm.

We know that many queries are syntactically naive and limited in scope [8-12]. It is also evident from the literature that the design of information interfaces affects query formation [13, 14], and that well-designed interfaces can help even those users who have poorly defined information needs [7, 15]. Library catalogues, however, have long stymied users in their efforts to find information [16, 17], as have research databases [18]. Consequently experienced academics and new students alike typically

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 260–271, 2011.

select Google and other internet search engines as their first choice of information resource [18, 19]. Despite choosing Google first, information seekers (at least in an academic setting) value libraries and library information resources [2, 20]. We also know from our earlier work that users of academic libraries perceive and use different types of library information (for example books and scholarly articles) differently, and are somewhat resistant to new trends to combine these resources into a single search interface [21].

While we do know that users use Google but value library resources, the literature gives little insight into the differences between the queries users enter into different kinds of information resources. Similarly, we could find no work that investigated whether any differences are due to the interfaces of such resources, or the due to the demonstrably different needs and expectations of searchers when looking for different kinds of information.

This paper reports an exploratory study of the queries entered into three different information sources via the same interface. This interface is on the homepage of an academic library website and records the initial queries users make of the library catalogue, EBSCOHost (a large library database), Gale (a smaller, more local database) and Google Scholar. It is our hope that by examining the content of these searches, we can learn about the differences in users' perceptions of these resources.

We commence by reviewing the background literature, followed by a description of the methodology employed in our study. The study's findings are then presented and discussed, leading to our conclusions and suggested avenues for future work.

2 Background

There is a long history of using search log analyses to understand user behavior in digital libraries [9, 10], physical libraries [12, 17], online research databases [11, 22] and on the web in general [8, 23]. These analyses have shown that users persistently accept interface defaults and use queries which average somewhere between two and three words long. This is true even where information seekers can reasonably be expected to have complex information needs. For example, a detailed study of a large academic article database shows similar query patterns by all but a very few users [22]. Not only do users accept defaults and use few words, but even in computer science databases (where searchers can be expected to be reasonably proficient with Boolean logic) Boolean operators and search limiters are not used (and where they are used, they are often used incorrectly) [10, 17].

The design of specific types of systems seems to be one factor upon user behavior. Studies of library users show they find library search services hard to use [2, 24], and that finding journal articles is particularly difficult [25]. One investigation of university students' academic library searching revealed that only library science students were prepared to search a range of resources or use advanced searching techniques; all other users preferred basic search [18]. Similarly, given the choice, most users of an academic library prefer federated searching to searching individual library databases, even though they know that they get higher quality results by the latter strategy [25]. Finally, library users strongly associate libraries with books. Even though, in academic libraries at least, usage of electronic information resources far

outstrips book loans [26], libraries are strongly associated with "books d'uh" [2] in users' minds. This causes problems when libraries ask users about their information seeking experiences, as most users report their experiences with books (c.f. [27]), depriving libraries of patrons' experience with what might be termed 'the online library'.

Earlier work by the authors in this area [21] asked users to comment specifically on their experiences with a wide range of information types, focusing particularly on books and scholarly journal articles. The results of this preparatory work suggest that academic library users, at least, view different types of library resource differently, and that books and scholarly articles in particular are used in different ways. Similarly, different academic disciplines use resource types differently, some focusing more heavily on journals, others on books [28, 29].

In summary, we can confidently generalize about library users' searching behavior: they do short, simple searches; they use a wide range of electronic resources; and they find library resources hard to use. The literature shows that books and scholarly articles are used in different contexts and to meet different information needs. What is unclear is whether users' search strategies change with resource type, and if so, whether these changes reflect something about the information sought, or are caused by the search interfaces. We set out to examine these significant questions through the study we now report.

3 Methodology

As noted above, research has faced the challenge that different types of resources are typically accessed through different interfaces. Integrated interfaces, if common in practice, have not received scrupulous scientific study. We exploited an integrated search interface that gives access to four different catalogs and its query log data.

We first describe the search interface in Section 3.1, and in Section 3.2 we report the analysis methods applied to the searches performed using the interface.

3.1 Search the Library

The Swinburne Library homepage has a single search box with radio buttons allowing users to pre-select among a range of information sources (see Figure 1).

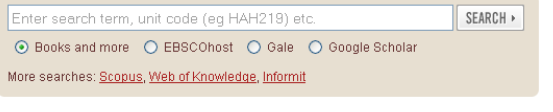


Fig. 1. The Library search interface

The default search is of the library catalogue ('Books and more') that includes physical books, ebooks, and a small number of articles made available through a course readings system. Three other resources can be searched directly: EBSCO, a large aggregator of journal content, Gale, a smaller journal content aggregator and

Google Scholar. Upon executing a search, this search box transfers the searcher to the native interface of the resource they searched, and as such the searches captured by this box (and analyzed in this study) are all initial searches rather than search refinements. We cannot tell from the logs alone how many results each search returned, though number of results could be approximated by repeating the search.

The predominant association of this interface is with the library: it is located on the library home page, and the default search target is the library catalog, neither of which are first choices for information seekers. While this association may affect the type of searcher using this interface, the information resource chosen can reasonably be assumed to be responsible for any differences in search strategy.

3.2 Analysis

The searches analyzed in this study were from two randomly selected weekdays, each during the second half of an academic term so as to represent practiced searching. A brief analysis of further days was conducted, but as they were not significantly different from the two days presented here, and as this study was intended to be exploratory, no further investigation of these days was undertaken.

The searches from the two selected days (3743 in total) were first analyzed in their entirety for query length and distribution over information sources. As it was neither realistic nor worthwhile in an exploratory study to manually examine these searches for content, following the gross analysis, a selection of 100 searches per day was taken from each of the three most used resources (the catalog, EBSCO, and Google Scholar—more on this in Section 4.1), for a total of 600 searches. Each query was then examined with respect to a number of factors: whether the target was a known item, the metadata used, what kind of target item chosen, whether the search used advanced search operators or contained typographical errors, and how successful the search was.

4 Results

In the two days examined as part of this study, users performed some 3743 searches encompassing 14648 search terms. 13 searches were empty and thus were excluded from further analysis. This section will present an analysis of query form, based on analyzing the entire dataset (Section 4.1) and a more detailed analysis of query content based on the sample data (see Section 4.2)

4.1 Query Form

While we could not examine all 3730 non-empty searches manually in this study, some automated analysis was possible, and interesting results emerged around the information sources users selected and the number of search terms per query.

4.1.1 Search Source

Of the searches analyzed, some 2277 used the library catalogue, meaning a full 1453 (38%) of searchers selected another target. This is a surprising finding, given that searchers typically change search defaults in fewer than 5% of cases [9, 17]. Because

neither of the sample days was near the beginning of semester, it is likely we are seeing practiced searching, with searchers preselecting an appropriate information resource for their needs.

Table 1. Number of words in each query

	Count	Mean	Median	Mode	Largest query
Library	2277	3.45	3	2	25
EBSCO	1104	4.58	3	2	26
Gale	56	4.55	3	2	16
Google Scholar	293	5.03	4	2	30
Total	3730	3.93	3	2	30

4.1.2 Search Terms

The number of terms in each query was counted and analyzed. The mean number of terms for all searches was 3.93, with Google Scholar having the highest mean (5.03) and the library catalogue the lowest (3.45) (see Table 1). Compared to the various studies reported earlier in this paper, these means are unusually high, and there are few searches with two or fewer words in comparison to other information interfaces.

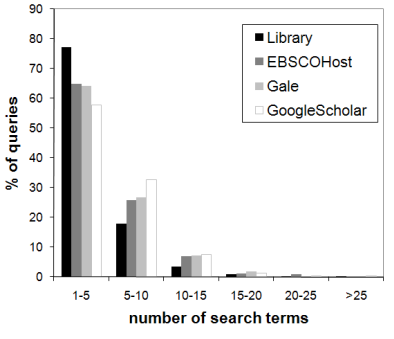


Fig. 2. Number of search terms by information source

The difference between the median number of searches and the mean in all cases suggests that there is a long tail of large searches (see Figure 2). Upon examination of the queries, one reason for this long tail emerges: searchers are typing or pasting whole citations as the query (e.g. *"Morsh, Joseph. (1930). Development of right-handed skill in the left-handed child. Child Development, 1(4), p311").* This behavior is a relatively new phenomenon (a citation search was used as an example in [17] but we could find no earlier reference to this behaviour) but may change the face of typical searcher behavior in the future.

Finally, it should be noted that while the statistics for each search source look broadly similar, the variance between information sources was statistically significant in a one way ANOVA comparison ($F(3, 3726)=44.02$, $p<0.01$).

4.2 Query Content

A randomly selected sample of 100 queries per day was taken from searches of EBSCO, Google Scholar, and the catalogue; these 600 searches were examined manually for metadata content, known item searching, typographical errors, advanced search techniques and search success. Gale was excluded from this phase of the analysis due to its small number of queries (56 in total).

4.2.1 Query Metadata

Searchers used a range of metadata, but by far the most popular were title, author and keyword. The distinction between keyword and title searches was made on the basis of matching: queries which matched three or more words in sequence in a resulting work title or any neologism in a title (e.g. "musicophilia"), or an exact title match of fewer than three words were deemed to be title searches (see Table 2). Title matches were exact in approximately 75% of all cases.

Table 2. Metadada types used in each search source

	Library	EBSCO	Scholar
Title	73	48	60
Author	33	19	19
Date	5	6	0
Publisher	9	1	13
Source	0	2	5
Edition/ issue	3	2	5
Pages	0	1	3
Keywords	109	133	129

Keyword searches (which can be assumed to be approximately equivalent to subject or topic searching) appeared in a variety of formats: some were traditional unconnected keyword searches (e.g. "*aviation fatigue*") some were phrases relating to users' information needs (e.g. "*differences in depth perception in 2d and 3d*") and in four cases (two each in Google Scholar and EBSCO), keyword searches were questions (e.g. "*how many international students have health insurance*"). The use of phrases or natural language queries occurred in about 15% of keyword searches in EBSCO and Google Scholar, and in 10% of searches of the library catalogue. This disparity is unsurprising, given that natural language searching is much more likely to be successful in online resources than in the catalogue.

When we examine the frequently used metadata types (metadata types seen in more than 5% of searches, i.e. title, author, date and keywords) there are differences in the frequency with which each metadata type is used between information resources. The differences between Google Scholar and the catalogue and EBBSCO and the catalogue are significant (χ^2=8.546, df=3, p=0.036 and χ^2=11.770, df=3, p=0.008 respectively). The queries directed at Google Scholar and EBSCO were broadly similar; there are a number of possible reasons for this including nature of the content searched (Scholar and EBSCO both search online article-based content) and expertise of searchers (users of these information sources have demonstrated some expertise by selecting a non-default option).

A small but not insignificant number of searches contained more than one type of metadata; some were as simple as combining author and title metadata while some users entered whole citations. Citation markers (such as the words 'et. al', and 'eds', and enclosing dates in brackets—which are semantically empty and thus cannot improve search success) were seen in 1.5% of searches in EBSCO, 3% in the library catalog, and 4% in Google Scholar. There was no significant difference between search sources in the number of metadata types used, and in all sources a greater number of metadata types is strongly correlated with a greater number of search terms (as we would expect).

Finally, the library catalogue saw the entry of a specific metadata type not seen in the other sources: In 3% of cases catalogue searches were for course-related material using an alphanumeric course code. As this type of search was not seen in the other interfaces, and it is relevant only to the catalog, this further reinforces the likelihood that users are making intentional choices about information resource selection.

4.2.2 Known Item Searching

Queries were considered to be for a known item if they used title searching, or if a combination of other metadata (such as author and date) identified an individual document. Non-title searches accounted for about 8% of all known-item searches, as per Table 3 below. There were significant differences (χ^2=8.998, df=2, p=0.011) between the three resources with respect to the number of known item searches,

Table 3. Known vs unknown item searches

	Library	EBSCO	Scholar
Known items	109	53	61
Non-title known items	11	7	4

Known item searches were classified by item type: by definition known item searches were for specific and identifiable texts, and each known-item query was classified as to the type of material sought. We discovered users were looking for a range of materials including books, articles and DVDs (see Table 4). It is interesting to note that (as we would hope, though the literature gives us no particular reason to expect this) known item searches of the catalogue are largely for books, while in EBSCO and Google Scholar they are mostly for articles. Scholar does show a number of book-related searches; however given that it returns relevant results from Google Books this strategy is likely to be successful.

Table 4. Item type sought

	Library	EBSCO	Scholar
Book	68	7	16
Article	8	41	43
Journal	2	3	0
Database	0	1	0
DVD	3	0	0
Other	2	1	2

4.2.3 Advanced Searching

As in other studies (e.g. [9-11, 17]) we found that only a small number of searches (4.5% across all resources) included advanced search techniques—13 (7.5%) each in EBSCO and Google Scholar, and 2 (1%) in the catalogue. The advanced catalogue searches were both well formed, but 4 in Google Scholar and 9 in EBSCO contained errors, echoing findings of other studies [9, 10, 17]. It is possible that more search modifiers may be observed if the interface captured search refinements, but the literature suggests that users are unlikely to use advanced query formulations even when refining searches.

4.2.4 Unsuccessful Searches

A search could be unsuccessful in one of three ways: 1) no results were returned; 2) the search was for a known item which did not appear in the top 5 results (bad results); 3) acceptable results only appeared with a 'did you mean' search modification. There were slight variations between interfaces in the number of unsuccessful searches (Table 5).

Table 5. Unsuccessful searches by information source

	Library	EBSCO	Scholar
Total	42	32	21
No results	13	10	5
Bad results	22	7	1
Did you mean	7	25	15

It is evident from these results that the catalogue is the least 'forgiving' system; this is to be expected as it indexes only relatively limited content (which may also help to explain why users consistently find catalogues unusable [6, 8]: catalogues simply do not index the material they want, and do not communicate this clearly). Google Scholar is the most 'forgiving' with the fewest failed searches and a minimum of searches with no or bad results.

Searches failed for a number of reasons. Typographical errors contributed to problems with search in about half of all search failures not all of which were corrected by "did you mean" functionality.

Another cause of failed searching (3 cases in the catalogue and 2 in EBSCO) was including an entire (unedited) citation in the search. The only search source that performed at all well when given so much information was Google Scholar; EBSCO and the catalogue both routinely failed to return any results on such searches.

A further cause of failed searching was searching in a system that did not index the kind of content sought (e.g. looking for books in EBSCO). While there was some cross-indexing, in all three systems some users failed to find what they were looking for because of the selected information source; there were 11(1.8% of all searches) instances in total —5 (2.5%) each in EBSCO and the catalog, and 1(0.5%) in Google Scholar. The low rate of failure in this way in Google Scholar is largely due cross-indexing of Google Books.

Finally, in six instances (3%) the library catalogue failed to return results because the library did not have a copy of the item users sought.

5 Implications for Digital Libraries

Many of our results mirrored those of other studies [9, 10, 17]: users still do not use advanced search techniques and still make a number of typographical errors. Digital library interfaces that expect users to perform complex queries without error are still likely to fail their users.

There were some behaviors in evidence, however, that demonstrate new patterns of behavior: e.g. users entered citations as query terms relatively frequently, increasing the average number of search terms, and often resulting in search failure. Digital library interfaces would do well to facilitate searching with entire citations.

We can reasonably assume that even those users searching the default information source (i.e. the library catalogue) generally expected to be searching for books; libraries are strongly associated with books [2] and only 2.5% of all catalogue searches examined in this study failed because the user was looking for articles or other resources more frequently indexed elsewhere. Unusually, the users in this study selected non-default search options 38% of the time—this is a dramatic contrast to other log analyses (for example [9, 17]) which show this behavior occurring only rarely. This suggests that the searchers in our study understood the differences between information sources and were intentional in their searching, a result previously only seen in topic experts and expert searchers [19, 29-31]. It is not clear from this study why users were so knowledgeable about their options, but it suggests that differentiation of content in an initial search can be valuable to users.

The fact that users in this study did differentiate their search strategies based on the target information resource demonstrates a level of intention in their search strategy not previously observed outside of experts in either domain or searching [19, 30, 31], and suggests that search strategies which have, in the past, looked naïve may be merely economical. This suggests that digital libraries should aim to support very simple searches rather than encouraging more complex formulations.

Despite this apparently high level of understanding of the options on the library homepage it is evident from usage statistics that far more academic library users access more electronic content than books [26] which suggests that many users are not searching for digital content directly from the library homepage. This means that the opportunity for digital libraries to be a single information point, which users would very much like [24, 25] has not yet been realized.

Finally, library resources are not a first choice of information source for many users [2, 24], so we can reasonably assume that those searching from a library website have entered the ‘active searching’ phase of information seeking [3]. Nonetheless, active searching doesn’t imply known-item searching, which accounted for just slightly more than half of all searches in the library catalogue, but less in other sources. The not inconsiderable number of non-known item searches implies users will browse search results quite heavily; digital library systems should take this into account and facilitate effective browsing.

6 Conclusions

This paper presents the results of an exploratory analysis of queries performed over four information systems: an academic library catalogue, two scholarly article databases (Gale

and EBSCO), and Google Scholar. The investigation was in two parts: a high level analysis of query structure over a large number of searches, and a more detailed analysis of query content based on sample searches of EBSCO, the catalogue and Google Scholar.

We found a surprisingly large number of searchers elected to search non-default information sources. Furthermore, there were differences in both the way users searched and in the results their searches returned, between different information sources. Longer queries were used in Google Scholar and EBSCO than in the catalog, but there were no differences in the number of metadata types queried across sources.

Similarly, queries of EBSCO and Google Scholar were more likely to be keyword searches than in the catalogue. Known item searches in Google Scholar and EBSCO targeted articles, whereas catalogue queries aimed at finding books. While citation searching was seen in all information sources, it was more common in Google Scholar where it was also more likely to be successful.

Given that we can discard interface effects, we must interpret these results to mean that users are intentionally varying their search strategies based on the information source searched. While the search strategies employed in each information source are similarly simple, they are different ant we must necessarily conclude that users goals and tactic vary specifically based on the target collection. This suggests that rather than being purely naïve, searchers tactic are rather parsimonious, a result previously only seen in domain experts and expert searchers.

We can (and should) take these results to show that digital libraries must continue to provide support for error recovery, allow users to preselect between collections and develop support for searching with unedited citations. More importantly, though, we must consider the possibility that users are more sophisticated in their search strategies than we had previously anticipated; the level of intention seen in this study is logical but previously undocumented.

7 Future work

Clearly there is scope for a more complete study to supplement this work: as this work was exploratory it analyzed only a relatively small number of queries, and an analysis of a larger sample to test the hypotheses about searching formed in this paper would be beneficial. Further, this work, like all log analyses, cannot unpack the motivations of searchers for behaving the way they do. The finding of non-default searching in this study is particularly striking, and it would be valuable to determine the drivers of this behavior in interviews or users studies.

References

1. Agosto, D.E., Hughes-Hassell, S.: People, places, and questions: An investigation of the everyday life information-seeking behaviors of urban young adults. Library & Inf. Science Research 27, 141–163 (2005)
2. De Rosa, C., Cantrell, J., Cellentani, D., Hawk, J., Jenkins, L., Wilson, A.: Perceptions of Libraries and Information Resources. In: OCLC, Dublin, Ohio, USA (2005)
3. McKenzie, P.J.: A model of information practices in accounts of everyday-life information seeking. J. Doc. 59, 19–40 (2003)

4. Marchionini, G.: Information Seeking in Electronic Environments, vol. 9. Cambridge University Press, Cambridge (1995)
5. Bates, M.J.: The design of browsing and berrypicking techniques for the online search interface. Online Information Review 13, 407–424 (1993)
6. McKay, D., Shukla, P., Hunt, R., Cunningham, S.J.: Enhanced browsing in digital libraries: three new approaches to browsing in Greenstone. JoDL 4, 283–297 (2004)
7. Nordlie, R.: "User Revealment"— A Comparison of Initial Queries and Ensuing Question Development in Online Searching and in Human Reference Interactions. In: SIGIR 1999, pp. 11–18. ACM Press, Berkeley (1999)
8. Jansen, B.J., Spink, A.: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Inform. Process. Manag. 42, 248–263 (2006)
9. Jones, S., Cunningham, S.J., McNab, R.: An Analysis of Usage of a Digital Library. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 261–277. Springer, Heidelberg (1998)
10. Mahoui, M., Cunningham, S.J.: Search Behavior in a Research-Oriented Digital Library. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, pp. 13–24. Springer, Heidelberg (2001)
11. Nicholas, D., Huntington, P., Jamali, H.R., Tenopir, C.: Finding Information in (Very Large) Digital Libraries: A Deep Log Approach to Determining Differences in Use According to Method of Access. J. Acad. Libr. 32, 119–126 (2006)
12. Wallace, P.M.: How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. RQ 33, 239–253 (1993)
13. Gerwe, P., Viles, C.L.: User effort in query construction and interface selection. In: DL 2000. ACM, San Antonio (2000)
14. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: SIGIR 1997. ACM, Amsterdam (2007)
15. Resnick, M.L., Vaughan, M.W.: Best practices and future visions for search user interfaces. JASIST 57, 781–787 (2006)
16. Cooper, M.D.: Usage patterns of a web-based library catalog. JASIST 52, 137–148 (2001)
17. Lau, E.P., Goh, D.H.-L.: In search of query patterns: A case study of a university OPAC. Inform. Process. Manag. 42, 1316–1329 (2006)
18. Griffiths, J.R., Brophy, P.: Student Searching Behavior and the Web: Use of Academic Resources and Google. Library Trends 53, 539–554 (2005)
19. Buchanan, G., Cunningham, S.J., Blandford, A., Rimmer, J., Warwick, C.: 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)
20. Blandford, A., Rimmer, J., Warwick, C.: Experiences of the Library in the Digital Age. In: 3rd International Conference on Cultural Convergence and Digital Technology. Foundation of the Hellenic World, Tavros (2006)
21. McKay, D.: Gotta keep 'em separated: Why the single search box may not be right for libraries. In: CHINZ 2011, ACM, Hamilton (2011)
22. Nicholas, D., Huntington, P., Jamali, H.R., Watkinson, A.: The Information Seeking Behaviour of the Users of Digital Scholarly Journals. Inform. Process. Manag. 42, 1345–1365 (2006)
23. Jansen, B.J., Spink, A., Saracevic, T.: Real Life Users and Real Needs: A Study and Analysis of User Queries on the Web. Inform. Process. Manag. 36, 207–227 (2000)
24. Fast, K.V., Campbell, D.G.: "I still like Google" University student perceptions of searching OPACs and the web. In: ASIST Proceedings, vol. 41, pp. 138–146 (2004)

25. Gore, G.: Undergraduates prefer federated searching to searching databases individually. In: EBLIP, vol. 3, pp. 61–63 (2008)
26. Martell, C.: The Elusive User: Changing Use Patterns in Academic Libraries 1995 to 2004. College & Research Libraries 68, 435–445 (2007)
27. Stelmaszewska, H., Blandford, A.: From physical to digital: a case study of computer scientists' behaviour in physical libraries. JoDL 4, 82–92 (2004)
28. George, C., Bright, A., Hurlbert, T., Linke, E.C., St. Clair, G., Stein, J.: Scholarly use of information: Graduate students' information seeking behavior. Inform. Res. 11 (2006)
29. Talja, S., Maula, H.: Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. J. Doc. 59, 673–691 (2003)
30. Marchionini, G., Dwiggins, S., Katz, A., Lin, X.: Information Seeking in Full-Text End-User-Oriented Search Systems: The Roles of Domain and Search Expertise. Library & Information Science Research 15, 35–69 (1993)
31. Vakkari, P.: Changes in Search Tactics and Relevance Judgements when Preparing a Research Proposal A Summary of the Findings of a Longitudinal Study. Inform. Retrieval 4, 295–310 (2001)

Understanding Documentary Practice: Lessons Learnt from the Text Encoding Initiative

Paul Scifleet[1] and Susan P. Williams[2]

[1] Discipline of Business Information Systems, Business School, University of Sydney, Sydney, Australia
[2] Institute for Information Systems Research, Universität Koblenz-Landau, Koblenz, Germany

How are definitions of content and the design of digital documents being determined in practice? In this paper the authors present the relationship between document encoder and document as the central unit of analysis in a framework for making sense of documentary practice at community, organisational and implementation levels. The paper presents the integrated findings from a global survey of document encoders participating in the Text Encoding Initiative, providing important insights into the characteristics of an emergent documentary practice. By focusing on documentation as a field of practice the paper reveals a rich and generative practice at play and provides valuable lessons for other complex metadata and markup initiatives.

1 Introduction

Markup language technologies are a significant part of the architecture for metadata initiatives in digital libraries and for sharing information resources over computer networks. However there is now a substantial body of evidence pointing to the challenges that the developers and users of large markup languages, such as Extensible Business Reporting Language (XBRL), HL7 (for health records) Legal-XML, and the Text Encoding Initiative (TEI), face in moving from community development to wider adoption and use [1] [5] [6] [7] [8]. Bringing a digital architecture to documents presents challenges of representation and meaning, organisational change, new work tasks and new business processes, yet to date research literature has discussed these issues either, in the most general of terms or, from the isolated experience of individual case reports. Although there are enormous volumes of digital documents being encoded across disparate fields of human endeavor it has been hard to gain a concrete understanding of practice beyond the individual accounts of practitioners. Much of the theoretical literature addressing the 'document problem' has done so broadly and at a conceptual level that often concludes by identifying the need for further research, both in the development of methodology and empirical observation [9].

The aim of our study is to extend existing research by providing an in depth investigation of documentary practices. In this paper we present the integrated findings from a large survey of markup language use within the Text Encoding Initiative's (TEI) community of practitioners that utilised three instruments for data

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 272–283, 2011.

collection: a questionnaire, the collection of text files for analysis and semi-structured interviews. The TEI is an international consortium of document encoders with a shared interest in the use of a document markup language (the TEI Guidelines) for the representation of all kinds of text in the humanities [10]. It is a longstanding initiative whose successes have informed projects undertaken by the World Wide Web Consortium and other markup languages like XBRL [11]. With more than 400 tags available for describing texts the TEI presents a flexible and diverse documentation system that is an ideal starting point for the investigation of documentary practices.

2 Research Design and Methodology

The theoretical orientation we bring to the study is informed by practice theory and asserts that it is the actual deliberations and decisions practitioners make in working with a markup language that has the most to tell us about document encoding [2] [10] [11]. Yet understanding documentary practice requires an investigation of much more than the rules, routines and regularities involved in 'marking up' a document. Document encoding, like classification, is a series of decisions about what should or should not be made explicit available about an artefact in its context. In this sense it is a form of documentation that joins with traditions of knowledge organisation and commentary in making judgments about the purpose of documents. Far from being static, axiomatic artefacts that reveal their own best description, documents are dynamic instruments where meanings, roles and context both structure and are structured by human understanding at different stages in the process of description.
Within markup language projects there are some obvious situational differences at play in practice, between the communities of interest who define the markup language, the organisations that adopt and adapt them and the individual document encoders who apply them and our study needs to account for these influences. Our interests are centred on the activity of *documenting*, which we see as a *relationship between documents and document encoders* who are involved in a process of decision making, from abstract notions about the types of documents that will be described and their expression (in rules and schema), through to their eventual construction and instantiation [11].

Our research objective is to better understand the structures and sets of structuring dispositions that are influencing practitioners in the material definition of documents at community, organisational and individual levels of interaction. The theory and methodology for this research have been detailed previously in two conference publications [10] [11] and are summarized briefly here. The study's participants are document encoders, working in universities, libraries, specialist digital humanities, centres, specialist documentation units and research institutes involved in significant document encoding projects using the TEI Guidelines. 32 document encoders from 12 countries participated in a global survey that was designed to gain a clearer understanding of their experiences of document encoding. Three interrelated data collection techniques were adopted in the study to support an interpretive approach to empirical observation that would account for the encoding relationship (between document and document encoder) at different levels of understanding:

1. The study commenced with a questionnaire based survey of practitioners that included both quantitative and qualitative (open-ended questions).
2. Participants contributed examples of texts they had encoded for in-depth automated analysis.
3. Semi-structured interviews were undertaken with practitioners to enrich our understanding of documentary practice and to gain deeper insights into the findings of the study.

Data collected in the questionnaire was tabulated and grouped according to three major categories derived from our analytical framework of the field of practice:

(i) community objectives and their expression in types of documents;
(ii) organisational objectives and their manifestation in document encoding;
(iii) implementation objectives and the resulting document instantiation

Data analysis commenced with the preparation of descriptive statistics and supporting narratives from the questionnaire. This allowed us to categorise and group findings first at a high level and then explore relationships to identified topics, aspects and themes of significance within each category. The iterative approach allowed for the identification of emergent, unexpected and incongruous themes [12] [13] [14]. We next undertook an in-depth automated analysis of the markup occurring in the encoded texts to see what the documents themselves could tell us about the categories, topics and themes we identified as significant from the questionnaire (and later, through this process of iteration, through the interviews also), looking in particular for a correspondence of relationships: e.g. do practitioners with a different professional background approach encoding differently?

By taking this approach to thematic analysis we are engaging in a process of qualitative interpretation of markup language definition and use to discover what it can tell us about this field of practice. Undertaking the interviews provided the opportunity to gain a deeper understanding of practice by allowing participants to present their own explanations of the phenomena witnessed in our analysis. It allowed us to confirm and elaborate topics coming forward and allowed viewpoints unanticipated by our framework for analysis to emerge [14].

3 The Lessons Learnt

Collectively the findings (presented as 12 key themes in Table 1 over page) define an emergent documentary practice that is rich, complex and generative. It comprises information about the historical circumstances of encoding, the thoughts and actions of individuals, activities, procedures and processes. What results is an in depth description of a documentary practice where information design, information management and knowledge sharing are central concerns. Below we discuss our key findings with a focus on lessons learnt in areas ranging from, the alignment of projects to long term organisational goals; strategies for communication; policy development; changing organisational structures; and the emergence of uniform approaches to analysis and design in documentary projects.

Table 1. Summary of lessons learnt in practice

Theme	Characteristics (from analysis)	Summary of key lessons learned
1.Fit for purpose	Utility (encompassing suitability, flexibility & extensibility), limitations (of the document expression)	Practitioners evaluate markup languages in terms of potential uses over time, in a wide strategy for adoption that focuses on the flexibility and utility of the language in meeting organisational goals for documentation (usually beyond a single project requirement). Changing a community language to meet local needs is an accepted part of practice
2. ML as standard	International, data compatibility (interchange and sharing), generalisable, extensible, often mandated. Prominence of TEI, Internationally recognised, funding	Practitioners encapsulate many of the advantages of markup languages within the umbrella of 'XML as standard'. More importantly, social dimensions, like community recognition of a markup language as the standard bearer, are key factor in driving adoption
3. Participation	Provides: Guidelines (DTD, schema), tools, training materials, support, confidence, communication, feedback	Actively encouraging participation in the document community is essential for the ongoing development of a markup language. Levels of participation in the community correspond to practitioner confidence
4. Communication	Collaboration, community, learning, support	Practitioners place greater emphasis on the opportunity to communicate directly with each other this than they do on formal training programs for the development of practice
5. Collaboration	Content, funding, scope (type & form of collaboration), communication & confidence, outcomes	Collaboration is fundamental to practice. Sharing content is the main driver of collaboration. Social interaction through collaboration is playing a key role in structuring practice
6. Mission	Governance & autonomy, organisational profile, emergent structures for digital documentation	Practitioners see document encoding as aligned to the core goals an organisation. Many limitations on encoding within an organisation are implicit, they are constrained by existing structures and shortcomings in resourcing
7. Policy	Formal policy, the role of learning and experience, extending the TEI Guidelines	There has been an absence of formal policy associated with markup language projects. When in place in TEI projects it has provided an important framework for learning through experimentation
8. Funding	The funding mix, pumping priming & ongoing funding, what gets funded (practitioner view), how funding influences on encoding	The main source of funding for TEI projects has been external and institutional grants. Much of this funding has been for project start up, resulting in significant difficulties for enhancements and developments that would allow important goals to be achieved. Funding beyond start up is essential if benefits are to be realized
9. Work unit	Document focus, specialisation, organising the encoding	There is clear evidence of the development of specialist digital units with a document-centric focus in the organisations, yielding advantages as new structures and new work roles come into play
10. Encoding	Text ontology, levels of encoding variation (similarity and difference) pragmatism	Practitioners prioritise the encoding of what is observable in a text. Challenging questions for classification, representation and meaning are a constant part of this work, which is best suited to subject or metadata specialists
11. Guidance	The guidelines, support, tools, structures and templates	Practice has resulted in the in-house development of important materials to support text encoding and provide the basis for the further development of practice
12. Design process	Products, design, evaluation and review	There is evidence of important digital document design and construction methodologies in place, commencing with strategic evaluation, text analysis, design and implementation, end user analysis and post implementation reviews

4 Discussion of Survey Findings

4.1 Community – Expression

Our study's focus on the practice of TEI document encoders commences with their principal work tool, the document expression that is contained in the TEI Guidelines. *'The TEI'* is shorthand in the TEI community for the set of instructions and tools that comprise the markup language schema (the vocabulary and its architecture), descriptions of each element (tag) and supporting documentation for their use.

Fit for purpose: our first theme gives voice to the significant evaluation that practitioners go through as they make determinations about adopting and working with the TEI. When asked: *What kinds of text do you consider the TEI most appropriate for?* The survey participants' view of the TEI seems to corresponds with the Consortium's own published statements about the kinds of text the TEI is suitable for; i.e. all types of text in the humanities (69%) prose (63%), drama (63%), manuscripts (63%), poetry (63%), verse (63%), transcripts (63%), guides and indexes (50% for each). More revealing though is the diversity evident in the 34% of responses that selected 'other' texts as a category and the long tail of documents this constituted in their replies: journals & serials, born digital materials & websites, correspondence, dictionaries, legal documents, lexical materials, emblems, steles & rubbings, religious texts. It is clear that practitioners are not simply applying a 'bolt-on' template that fits all texts, the TEI is more accurately a complex document system comprising over 400 descriptive elements (tags) that practitioners can choose from to assemble the sets of tags that meet their needs. Creating a TEI text commences with decisions made about the documents that will be constructed from this system. We analysed 27 batches of encoded text and only two batches (using a predefined set, known as TEIXLITE) shared the exact same elements. The remaining 25 batches of documents were all using document type definitions that were different from each other. Notably, practitioners rarely chose their TEI set solely on immediate requirements. Each of the 15 participants who took part in the interviews for the study stressed the importance of the utility and flexibility of the TEI in meeting both current and ongoing documentary tasks. The TEI Consortium supports this by providing an online suite of tools (the Pizza Chef and Roma) that practitioners use to select the mix of elements (pizza toppings) they need [15]. Even so, adopting a community schema to meet local requirements is bound to bring constraints. Eleven of 15 interviewees indicated that they are constantly evaluating the TEI and wary of "shoehorning" content to a model that doesn't fit their needs (B021)[1]. Three participants spoke about the importance of commencing assessment of text encoding requirements at a higher level of abstraction than the TEI's model of a text as important. They considered that there were other ways of representing the same texts (using different markup languages) and that limitations in the Consortium's view needed to be evaluated: *"…if one is going to spend a significant amount of time and effort to morph the TEI into being useful for the encoding you're trying to perform, and there's another system out there that is well suited to the type of encoding you would like to perform, you're probably better to use that system."* (B028)

[1] All direct references to survey participants in the paper use assigned anonymous identifiers, e.g. (B021). All quotes from interviewees are included in italics.

Fit for purpose as it emerges here is not about fitting a model or matching a markup language to the requirements of specific projects, it is a general quality of fitness to the strategic objectives for documentation over time. Importantly, the evidence does show that document encoders are evaluating the advantages of the XML standard (e.g. suitability, flexibility and extensibility) as part of this and they are prepared to work with the constraints they identify to obtain the benefits of standardisation. However, fitness is a balancing act between the benefits and the limitations that a markup language brings.

Markup language as standard bearer: for some participants the use of TEI has been mandated precisely because it is XML conformant: *"…one of the mandates was that whatever we did we should follow best practices for archival storage of documents in digital form. And that lead us to the TEI model, since this is the de facto standard for the markup of text."* (B020) Perhaps more important in our study however, has been the significance encoders place on the TEI as *the standard bearer*. Rather than focusing their evaluation on the technical advantages of XML, practitioners are evaluating the emblematic and social value (both organisational and cultural) of the TEI standard: *"…when we first started up there was a lot of people using TEI in the library, in digital humanities institutes and so on. So I guess it was peer pressure in the widest sense"* (B020). Eight of the interviewees identified *TEI as a standard bearer* as a significant reason for choosing to work with the TEI, whether that be for recognition, because its preeminence improves funding opportunities or, from 'peer pressure in the widest sense.' Emergent in this has been the significance placed on the profile of a markup language that is compliant with technical standards and community expectations. A new insight here is that the latter is equally important in driving adoption. In this light, the normative influence of standardisation resides not only in the structuring dimensions of a technical specification but also in the orthodoxy of the community.

Participation, Communication, Collaboration: Three social dimensions of practice: *participation* (in the Consortium), *communication* (with each other) and *collaboration* (with other practitioners and institutions within the field) entwine to shape knowledge sharing among practitioners. The TEI Consortium encourages participation from its members and this has contributed to the long-term success of the markup language by ensuring practitioner feedback actively shapes iterative development and provides a sense of ownership of the markup language. So much so that two interviewees described their relationship with the Consortium as *their most important collaboration in document encoding*. Twenty of the survey respondents were active members of the TEI, 16 had at least one active role (committee or special interest group), while 10 were active in other markup language and metadata initiatives. While this brings forward an obvious bias in the point of view of our participants, it also underscores the significance of engagement in practice. As other markup languages come and go within their fields of specialisation, there must be lessons that can be learned from this. Our analysis shows that document encoders actively participating in the Consortium through committees and events had a much higher level of confidence about their work than practitioners taking a more passive role in the community. Ten of our interviewees emphasised the importance of the Consortium in developing their confidence. However, five survey participants indicated that they felt they lacked the expertise necessary to participate: *"I feel like I*

am a user of the TEI Guidelines but not somebody… who's in a postion to change them or even discuss them with others. Because I feel I'm still not a programmer, …I still think you need a lot of extra knowledge to contribute to the Consortium. But as a user I am very happy with the fact that it exists, and I can use it in the way that we think is okay." (B021) Understanding how confidence is developed tells us something about the way design choices are brought to bear across different encoding projects through participation and communication within the community. Importantly, practitioners place greater emphasis on communication and interaction through, the TEI LISTSERV, example files from recognized lead projects, and emails than they do on formal training programs. Paramount among these is 'the list': *"I think the TEI list, whilst I don't always agree with everything that's said on the list, nevertheless, it's a really important part of the TEI.* (B021) One participant described communication as one of the greatest benefits of collaboration, *"…even if the collaboration doesn't directly benefit us. Being able to talk to others and seeing what they're thinking about, that's useful."* (B028). Collaboration is an important characteristic of document encoding projects overall. 78% of participants were involved in collaborative projects. Sharing content is the main reason for collaborating. Most collaborative projects were either collection based (19 corpus or papers) or subject and theme based (16 projects) initiatives. Collaboration provides the opportunity to unify disparate content and publish materials that might otherwise not be published. However it also shapes the task as participants adapt to requirements of collaborations that can influence both what is selected for encoding and how it is encoded (depth, detail and tag choice), even how it is published: *"…in terms of putting our own texts into another project, we need to be very careful with our metadata and make sure that everything is going to be reciprocal."* (B012); *"Our use of text encoding fuels their website.* (B021); Collaboration both enables and constrains practice, however it is clear that the benefits extend beyond the opportunities available through funding and content sharing to the important role that social interaction plays in structuring practice.

4.2 Organisation – Manifestation

Both Nunberg (1996) and Frohman (1994) hypothesise that much of the capacity of documents to inform will depend on the institutional arrangements for their production [9] [15]. Our inquiry aimed to learn more about how institutional structures may manifest in the encoding activity. The findings from our analysis identify the kinds of structures organisations adopt to facilitate digital document production and the new mix of professional requirements that are needed to achieve this. Rather than being a simple continuation of the traditions of humanities scholarship, librarianship and publishing, there appears to be a new practice emergent that is more than a sum total of its parts.

Mission, policy and funding: A challenge for the study has been reconciling the perception of autonomy and absence of institutional influence reported by participants with evidence that, at times, demonstrated the strong effect of institutional arrangements. 73% of the study's participants rated their autonomy in the encoding choices they make as either *highly autonomous* or *completely autonomous*. In some respects the perception of autonomy is not surprising, text encoding takes place within

a scholarly community that values intellectual autonomy. Yet, in one form or another, the survey participants discussed the constraints they experience through institutional arrangements: 'Market demands', limits on funding and editorial management were all brought forward as effecting choices about which texts were selected for encoding and how they were encoded (i.e. on the textual outcomes). *All* of the interviewees discussed a level of service orientation in their work that indicates a shift from traditional humanities scholarship influenced by a microcosm of new interests associated directly with the activity of encoding. This includes the production of new digital information resources designed in some way to service the needs of others (scholars, students, a largely unknown public) or, meeting the demands of a publisher's database. As one participant observed, "*…under these conditions many constraints are not noticed until they are breached*." We consider the contradiction between the perceived autonomy of practitioners and the impact of organisational arrangements on document encoding choices, as one that is best explained by the subtlety of influence in structures often only evidenced in the general context of organisational doxa. Rather than being direct (e.g. the prescription of document templates), institutional arrangements were most often indirect, through factors like the provision of in-kind support or, limitations on the availability of resources.

61% of participants indicated that their organisation had no formal policy in place for document encoding at the commencement of their projects. A topic analysis was undertaken to evaluate similarities among policy documents from those respondents who did have them. The policies that were in place tended to present guidelines inline with objectives for scholarship but they did not prescribe or constrain specific aspects of using a markup language. Those aspects that were prescriptive generally directed projects towards conformance with XML standards as best practice. The most significant finding coming forward from policies is that experimentation has often been part of their ethos, *"…that pushes our learning curve in a consistent direction; start-up, pilot completion based on proof of concept, expansion of project."* (B029). By experimenting, the projects have played a significant role in constructing a shared practice. Stated formally in one policy as: *"To document and analyse the success or failure of each of these plans, and to publish that information for analysis for the benefit of other academic publishers, policy makers and authors."* (B020) The findings demonstrate how the shape of documentary practice begins to take form, is shared with others and reproduced through models of experimentation and learning.

Participant responses indicate that the greatest cost constraint on a project is the cost of doing the encoding. Funds available for projects impact not only the selection of texts and the richness of encoding that can be done but the kind of digital outputs that will result. To date the majority output of text encoding for public use has been in the form of HTML files. This, unfortunately, misrepresents the work behind the scenes where the encoding of texts has been rich in detail. While grant funding has often been available to projects at start-up, finding revenue for ongoing project development is a major concern. Almost every participant in this study identified enhancements they could now bring to their projects. However, obtaining funding support for enhancements that utilise embedded semantic structures in novel ways or, investing in the improvement of encoding methodologies, is rarely a funding priority. While 29 of the survey's participants anticipated that document encoding would be an

ongoing activity in their organisation, moves to self-funding models (e.g. licensed databases) further reduce depth and variety in the encoding that can be done.

The Work Unit: The TEI's adoption as a model for digital library projects has raised some concerns for practice [16]. Scholars generally want to operate under a model that is (analytically) as flexible as possible while in libraries classification is often prescriptive, target consistency to maximise user access [16]. At one level this difference between two communities sharing one markup language seems to be born out in our study. While scholars rated their autonomy as *"highly"* or *"completely autonomous"*, all participants who worked in libraries rated their autonomy only as *"somewhat autonomous"* or '*not autonomous at all."* However through a careful analysis of the markup in all batches of texts submitted we could find no measurable difference in encoded documents that could be attributed to different areas of managerial responsibility (i.e. there were no differences in textual outcomes that could be attributed to academic publishers, libraries, faculty, or other) or, to professional background (the encoding of librarians was not distinguishable from academics or academic publishers). While differences do occur at project levels of description, the most significant pattern to emerge was the uniformity of tags used across texts within a project/work unit: indicating that once document encoders develop a model for encoding they tend to work with the same range of tags consistently across texts with only occasional variation to account for particularities of individual texts (this was usually an increase in tag frequency, i.e. the same tag used more often; rather than an increase in the range of tags used).

Rather than following traditional lines of professional difference, document encoding is a new activity and new structures are being implemented to support this. By grouping and categorising different project roles defined by the study participants we were able to develop a view of an organisational framework that is not simply utilising existing structures but establishing new approaches and new roles for organising, managing and resourcing digital document projects. For most projects, an unexpectedly complex picture of collaborative production emerges, involving project directors (usually from academic faculty, University libraries or the University press); IT specialists; content, metadata and markup specialists; electronic collections management and preservation specialists; editors; financial management; and specialist advisory roles including legal advice on copyright and intellectual property. These arrangements are usually localised within an institution: i.e. a mix of suitable skills is being developed in-house through practice to support a different way of doing things. This is an important profile of organisational arrangements that did not exist 25 years ago when the TEI commenced. Most notable are the new specialised units and functions in support of the activity with four humanities computing centres, two digital production units, specialist document centres, and faculty knowledge technology centres represented in this study. Awareness of the requirements for specialised arrangements, with new skill sets and organisational structures flags important challenges for the diffusion of digital document initiatives in other fields of endeavor such as health, law and business.

Within these emerging roles, the design decisions practitioners make are strongly influenced by their own culture, *as document encoders*. The task and office can seem to share some aspects in common with librarianship (texts are prioritised according to principles similar to a library's collection policy, both use classification systems) and

it is deeply embedded in a tradition of humanities scholarship where construction of scholia and the production of scholarly editions are longstanding, yet it is neither of these exactly. Document encoders bring their history and the experience of their traditions with them and it is helping to shape the activity, but it is taking its own shape in "practical mastery of the specific regularities" of the field and there are some distinguishing features associated with this [2]. Among these is the *document centric* focus that is driving practice. Encoders focus their design task more on their representation of the content of a document than they do on any other influence, with 47% of participants ranking 'the requirements of the text' itself as the greater criteria for content definition (with 29% identifying end user requirements, 9% institutional requirements and 15% community objectives). There is no simple or right answer to question of what constitutes 'the text itself. Documentation is an activity where both subjective and objective aspects of the act are real considerations in practice.

4.3 Implementation - Instantiation

Encoding: 92% of participants described their own level of encoding as providing for more than simple analysis of the text, with 50% describing the level of encoding they undertake as producing texts suitable for scholarly analysis. Typically this level of encoding provides for intellectual, linguistic and prosodic description of text and is broad enough to encompass tagging that is itself a significant commentary about the work. Document encoders describe undertaking this work as a 'balancing act' between, addressing the complex ontological issues of representation and meaning that arise when encoding, and the more pressing concern of completing encoding in a reasonable amount of time: *"..we've rationalized the choices in ways that we are comfortable with, but realizing that there are other possibilities, that other people may have made other choices given the materials and the possibilities for encoding them."* (B018)

In the case of the TEI, practitioners are most often prioritising representation of what is observable in a text, however this remains an activity with many nuances & variations. 53% of participants take advantage of the TEI Pizza Chef tool to modify their document type definitions (DTDs) at the outset, while 50% are modifying the DTD further in-house. 94% of participants in the study assign attributes and attribute values to the tags they use and this also results in different patterns of use. 47% indicated the adoption of a variety of different classification schemes (with the most common being Library of Congress Subject Headings) while some were creating their own inhouse classification schemes as part of activity of textual criticism.

Guidance and design process: Markup languages present a number of characteristics that we would anticipate have a normalizing effect. Categorisation is possible because an agreed set of terms and the conventions for their use are adopted. Yet the evidence from the survey consistently presents a practice that is localized to its documentary-subjects, and its environment. It can appear as though there are very few formal, shared analysis and design techniques in place to assist in practice, and this view was reinforced by some of our participants: *"We have done nothing which I think would fit the requirements of doing such analysis properly. None of us is trained in doing such analysis."* (B027) However, closer examination reveals that analysis and design techniques are in place with a significant amount of uniformity across

projects, although this fact may not yet be fully apprehended by the practitioners. Our interviews identified clear steps in digital document design that practitioners are undertaking across projects. In this paper we have demonstrated how practitioners commence by aligning their work to strategic goals and organisational objectives through an analysis of *fit for purpose*. Furthermore, all participants identified processes equivalent to text analysis, planning and designing markup solutions for implementation (in that order). Supporting this is a range of implementation methods and tools including in-house codebooks, guidelines and procedure manuals to guide encoding. Many practitioners are conscious of the need to evaluate end user needs and are undertaking significant post implementation reviews as a step in this direction. However, how user requirements should be constituted in the environment of TEI encoding remains challenging: *"What we do, is doing these things as we do them, and make them available for users and see how they react and take seriously the responses and criticisms and suggestions"* (B027). Although most practitioners may not yet be identifying (or accounting) for them in this way, we found that discussion with stakeholders, including other document encoders, informal feedback from users, journal reviews, and explaining the text system to others (at seminars and conferences) were all in play as part of the design process.

5 Implications and Conclusion

Important in this study has been the objective of investigating documentary practice as it is emerging in the TEI community so that we might arrive at an understanding of practice detailed enough to be useful to other document encoding and metadata initiatives. The study has not sought to identify and describe *the essential criteria for best practice*. Rather it has sought to understand the interactions between document encoders and their documents as they work with them in different contexts defined by community, organisational and implementation level concerns.

The themes and characteristics that have been tabled (Table 1) in this paper focus on those aspects of practice that commence a practical understanding that should be transferable to other domains. Information ranging from communication and policy strategies to funding models and project pathways is now available to support practitioner confidence in decision-making and we anticipate that this study's reports various reports will contribute in this way. It is likely that different and opposing characteristics will come forward through studies of other documentary communities. A field of practice is always a dynamic space that is constituted by the interaction of its elements. It must allow for the choices practitioners make through their *feel for the game* in each situation [2]. This observation is a critical aspect of this study that distinguishes it from other research where the document has been viewed simply as an inert artefact to be tagged. It contrasts sharply with the mechanistic determinism that often frames the discussion of markup languages. While we anticipate that different domains with different documents and different documentary needs will prioritise different characteristics, further research now has a starting point with a framework, themes and topics identified that will support that analysis.

References

1. Megginson, D.: Imperfect XML. Rants, Raves, Tips and Tricks from an insider. Addison Wesley, Upper Saddle River (2004)
2. Bourdieu, P.: Outline of a Theory of Practice. Cambridge University Press, Cambridge (1977)
3. Brown, S.: The Orlando Project: Origins and Aims. In: The Orlando Project: Humanities Computing in Conversation with Literary History, Association for Computing in the Humanities and the Association for Literary and Linguistics Computing, Kingston, Ontario (1997)
4. Päivärinta, T., Tyrväinen, P., et al.: Defining Organizational Document Metadata: A Case Beyond Standards. In: 10th European Conference on Information Systems, Gdańsk Poland (2002)
5. Debreceny, R., Gray, G.L.: The production and use of semantically rich accounting reports on the internet: XML and XBRL. International Journal of Accounting Information Systems 2, 47–74 (2003)
6. Wrightson, A.: Is it Possible to be Simple Without Being Stupid? Exploring the Semantics of Model-driven XML. Extreme Markup Languages
7. Mountain, D.: XML E-Contracts: Documents that Describe Themselves. International Journal of Law & Information Technology 11(3), 274–285 (2003)
8. Sperberg-MacQueen, C.M., Burnard, L.: Guidelines for Electronic text Encoding and Interchange, The TEI Consortium (2004), `http://www.tei-c.org/P4X/index.html` (accessed online 2010)
9. Frohmann, B.: Documentation Redux: Prolegomenon to (Another) Philosophy of Information. Library Trends 52(3), 387–407 (2004)
10. Scifleet, P., Williams, S.P.: Practice Theory & the Foundations of Digital Document Design. In: Proceedings of the 27th ACM International Conference on the Design of Communication, SIGDOC 2009, Bloomington, USA (October 7, 2009)
11. Scifleet, P., Williams, S.P.: Constructing Digital Documents: Emerging Themes in Documentary Practice. In: Proceedings of the 44th Hawaii International Conference on Systems Sciences HICSS-44, Hawaii, USA (January 7, 2011)
12. Bogdan, R.C., Bilken, S.K.: Qualitative Research for Education. Allyn and Bacon, Needham Heights, MA (1992)
13. Lofland, J., Snow, D.A., et al.: Analysing Social Settings. Wadsworth Thomson Learning, Belmont (2006)
14. Miles, M.B., Huberman, M.A.: Qualitative Data Analysis. In: An Expanded Sourcebook, Sage Publications Inc., Thousand Oaks (1994)
15. Nunberg, G.: Farewell to the Information Age. In: Nunberg, G. (ed.) The Future of the Book, pp. 103–138. University of California Press, Berkeley (1996)
16. Hockey, S.: History of Humanities Computing. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) A Companion to Digital Humanities, pp. 3–19. Blackwell Publishing, Malden (2004)

Linking FRBR Entities to LOD through Semantic Matching

Naimdjon Takhirov, Fabien Duchateau*, and Trond Aalberg

Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
{takhirov,fabiend,trondaal}@idi.ntnu.no

Abstract. In this paper, we present an approach to automatically link FRBR works identified in metadata to the corresponding entity in Linked Open Data resources. The main contribution is a basis for semantic enrichment and verification of works identified in existing metadata. Through experiments, we demonstrate that FRBR works can be identified in the LOD cloud, which provides a solid ground for further work.

1 Introduction

Metadata related to cultural items such as movies, books and music is a valuable resource that is currently exploited in many applications and services based on mashup and linked data. Semantic Web technologies can be used to expose and interpret the meaning of the data on the Web, publicly available API's enable third parties to develop innovative services for existing data, and new knowledge can be created by linking related and complementary data from different sources.

The use of conceptual domain models is an important part of this environment as they define the universe of discourse and facilitates the proper semantic integration of the information within a domain. The Functional Requirements for Bibliographic Records (FRBR) [10] is a conceptual model that increasingly is being recognized as the common domain model for cultural items, and one of the main challenges deals with the interpretation or conversion of existing data into FRBR-based representations. The proposed semi-automatic approaches [8,13,1] have been designed to fulfill this goal but they mainly focus on converting bibliographic records found in library catalogs.

The Linked Open Data (LOD) vision [2] and the increasing demand for semantic aware data has strengthened the interest in FRBR. In this paper, we present a solution for linking the FRBR *works* that can be identified in metadata to its corresponding LOD entity. The main motivation is to bridge the gap between metadata that mainly identify such entities through implicit descriptions and the explicit representation of these entities that we can find in LOD resources such as DBpedia and OpenCyc. The benefit of this solution is the ability to semantically enrich existing metadata with attributes and relationships discovered when

* The author carried out this work during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 284–295, 2011.

linked to LOD, but we will also argue that our approach can serve as a basis for verification purposes, specifically for tools which automatically convert legacy data into FRBR. We demonstrate that our approach is effective by performing a set of experiments with Amazon product data.

2 Preliminaries

2.1 Functional Requirements for Bibliographic Records (FRBR)

The FRBR is a conceptual model of the bibliographic universe published around a decade ago [10]. It models intellectual and artistic endeavor in multiple levels of abstraction: **work** (e.g. *The Two Towers* by J.R.R. Tolkien) , **expression** (e.g., *To Tårn*, a Norwegian translation of that work), **manifestation** (e.g., a paperback format of this expression published by Mariner Books in 2005), and **item** (e.g. a physical book). A **person (or corporate body)** in the FRBR model, is an individual responsible for the creation or realization of a work (e.g., as an author, an illustrator, a translator, etc.). Additionally, the FRBR model provides **a set of relationships** between entities beyond the basic relationships.

2.2 Linking Open Data

The Semantic Web is a technology to support the web of data (contrary to the current web of documents) by relying on semantic and linked data, models (e.g., RDF), query languages (e.g., SPARQL), inference system and applications. More specifically, **Linked Open Data** (LOD) encompasses a vision in which all data is globally accessible and interconnected, thus making it more valuable. All LOD entities such as subjects or properties, are identified by a unique Uniform Resource Identifiers (URI). The LOD cloud refers to interconnected data sources, such as DBpedia, Freebase or OpenCyc, which can be seen as the foundation of the LOD vision. With the emergence of this initiative, an increasing number of data sets is published as linked data. The basic principles of publishing on the LOD cloud is the use of RDF as a data model and RDF links to interlink data from diverse data sources. The primary motivation for publishing data in a LOD cloud is it provides a basis for semantic reuse and integration of data from diverse sources. To reach this goal, data should be represented in a well-defined structure [2].

3 Related Work

Our approach consists in linking a FRBR work to its corresponding LOD entity, and it lies in the intersection of two domains. The former is **entity search**, since we want to discover equivalent entities based on their information. However, one of the entities we intend to match is a semantic entity in the LOD cloud, which deals with **entity ranking**. The rest of this section provides more details about these two research domains.

The *entity search* problem, also known as *record linkage* or *entity resolution*, is a crucial task for data integration or data cleaning [7]. It mainly aims at identifying entities (objects or data instances) which represent the same real-world entity. Contrary to existing approaches, which are designed to match entities represented in a relational framework [11], we apply entity search to RDF entities. Besides, most of them are based on machine learning techniques, and require training data. Another major difference deals with the quality of the data sources: in our context, we can assume that the data from the LOD cloud does not contain many errors for a given entity.

On the Web, a similar task, called *entity ranking*, involves the discovery of an entity's main page, contrary to traditional search engines which propose documents mentioning a given entity. This task has been extensively studied and two initiatives have an entity ranking track arranged every year: *Initiative for the Evaluation of XML Retrieval* (INEX) [9] and *Text Retrieval Conference* (TREC) [17]. Most approaches which take part in these tracks are either based on information retrieval or semantic web [16,14]. The main difference between our work and entity ranking is the availability of information. In our context, the type of the searched entity is not always specified, or with a broader topic. Conversely, the work that we want to match to a LOD entity can include useful information such as creator, year, or categories.

4 Matching FRBR Works to LOD

Linking FRBR entities to the LOD cloud is a solution with many benefits. First, it could enable the automatic enrichment of FRBR entities discovered in existing metadata with additional attributes and relationships. For instance, we may discover the relationship between a book and the screenplay that is based on the same novel by looking up the work in the LOD cloud. Secondly, the LOD cloud can be used to verify or guide the FRBR-based interpretation of existing information provided about the product, which can be misleading and ambiguous when interpreting the intellectual aspects due to a lack of semantics. For instance, when using titles and authors to identify works there can be a large number of false positives if there are many translations or adaptations of the same work. The LOD cloud can be used to verify the proper works or to single out the work entities that are of main interest to end users. In the rest of this section, we explain how we discover a relationship between a FRBR work and a LOD entity.

4.1 The Problem

We have a set of works $\mathcal{W}$ and a set of LOD entities $\mathcal{L}$. Note that the LOD entities are linked to other entities by relationships, but we do not need this feature at this stage. All works and entities have a set of attributes. Considering a work $w \in \mathcal{W}$ and a LOD entity $l \in \mathcal{L}$, we note $\mathcal{F}$ the set of attributes shared by w and l. To assess a degree of similarity between w and l, we compute similarity values between their shared attributes. For an attribute $f \in \mathcal{F}$ shared by w and l, a similarity function is defined as follows:

$$sim_f(w, l) \rightarrow [0, 1]$$

The similarity function returns values between 0 and 1, 0 indicating attribute f of w and l is completely dissimilar and 1 if the attribute f is identical for both w and l. The amount of data available on LOD is very large, and even the querying of only one data source can be time consuming. Thus, the goal is first to reduce the search space by obtaining a subset of LOD entities, a process called *Blocking*. Then, we can apply fine-grained *matching* techniques on these entities to compute their degree of similarity with the given FRBR work. This process outputs a ranking for these LOD entities according to their similarity degree. Figure 1 sums up our approach for matching FRBR to LOD. As a running example, we use a work entitled *The fellowship of the ring (LOTR)*. It includes the following (incomplete) list of attributes: *novel* as type, *JRR Tolkien* as creator, *science fiction & fantasy* for categories and no creation date.

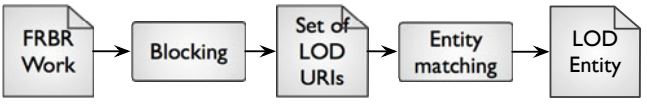


Fig. 1. Workflow of Entity Matching

4.2 Blocking

In entity matching, the large amount of entities implies to have a method for reducing the search space. For instance, if an entity has a title, a simple blocking method could be the matching of entities that share at least a common word in their titles. In our context, this blocking process is required for two reasons. First, we query remote services with their potential issues (e.g., network overload, query limitations). Secondly, the set of entities is very large: more than 3.4 millions entities for DBpedia[1], 12 millions for Freebase[2] and thousands of entities for OpenCyc[3], which are only a few of the data sources from the LOD cloud. As a consequence, we need to have a heuristic to retrieve only a subset of entities against which we apply matching techniques. The following techniques can be used to search for a LOD entity:

- Knowing or generating the correct URI of the entity, which cannot be applied in our context;
- Querying SPARQL endpoints;
- Querying a Lookup engine.

Both SPARQL and Lookup queries can return a set of LOD entities that match the search query. Thus, we reduce the search space by using these services, since

[1] http://dbpedia.org/, January 2011.
[2] http://www.freebase.com/, January 2011.
[3] http://www.opencyc.org/, January 2011.

they return an acceptable number of results (in our case usually between 0 and 200). In order to increase the probability of obtaining the correct entity in the search results, we need to build different queries based on the information contained in the work's attributes.

We have identified three interesting attributes of a work that can be used to generate a set of queries: title, creator and type. However, these attributes cannot be used directly in the query. They need to be transformed to remove extra information, to split creator's name, or to broaden a type. The idea is to create a set of query tokens for each of these attributes. More formally, we want to obtain three sets *titles*, *creators* and *types* containing query tokens such as:

$$titles \rightarrow \{title, normalized\_title\}$$
$$creators \rightarrow \{creator_1, ..., creator_k\}$$
$$types \rightarrow \{type, ext\_type_1, ..., ext\_type_m\}$$

The *titles* set contains the full title of the work, and a normalized title in which extra information (e.g., inside parenthesis) and useless grammatical words are removed. In other words, this normalized title only includes the most important words after a normalization process [6]. For the creators set, each creator's name is used as a query token. Finally, the *types* set contains the type of the work and its extensions. These extensions are hypernyms and synonyms from a predefined list (from Wordnet[4]), e.g., the type *novel* is extended with *print* and *book*.

Once we have produced the three sets with their query tokens, we can combine the query tokens to generate a query. Combining these tokens is required either to obtain more results or to disambiguate. For instance, the novel entitled *airport* only returns a list of airports if the type is not included in the query. So the idea is to perform all combinations of 1, 2 or 3 tokens, each token belonging to a different set, and use these combinations as queries. All results returned by each query are merged based on the unique entity URI. Note that if all individual tokens do not return any results, there is no need to send queries which include this combination. At the end of this blocking process, we obtain a set of LOD entities (represented by their URI) against which we apply refined matching techniques.

We have generated different queries for our example work dealing with *The fellowship of the ring (LOTR)*. Table 1 shows some of these queries and provides the number of results returned by a Lookup service. Here, we highlight the need for sending multiple queries. Even with a well-known artistic work such as *The fellowship of the rings*, the lookup did not return any results with the full title, hence the need to simplify this title. Similarly, a query including the normalized title and the type of the work did not provide any results, contrary to the normalized title combined with an extended type.

4.3 Entity Matching

After the blocking step, we obtain a normalized set of LOD entities, and we need to match them against our work. To fulfill this goal, we first identify which

[4] http://wordnet.princeton.edu, January 2011.

Table 1. A Subset of Generated Queries for our Work Example

Type of Query	Query	# Returned Entities
title	The fellowship of the ring (LOTR)	0
norm_ title	fellowship ring	5
title + creator	The fellowship of the ring (LOTR) JRR Tolkien	0
norm_ title + creator	fellowship ring JRR Tolkien	0
norm_ title + type	fellowship ring novel	0
norm_ title + ext_ type	fellowship ring book	1
norm_ title + ext_ type	fellowship ring print	0
creator + type	JRR Tolkien novel	0
creator + ext_ type	JRR Tolkien book	1

shared attributes can be matched, and then we describe the similarity functions applied to these attributes. A global similarity value between a work and a LOD entity is finally computed, and filters may be used to discard some of the matched entities.

Identifying Attributes. First, we have identified the most important attributes that we can use to compare a FRBR work and a LOD entity. Although these attributes depend on the data sources we have on both sides (work and entity), five attributes are at least very common:

1. Title. In our running example, the work title has the value "The fellowship of the ring (LOTR)";
2. Type of work/entity. For instance, the work type of *The fellowship of the ring (LOTR)* is "novel" while the type of the corresponding entity is "book";
3. Creator. All artistic works have one or more creators. "J.R.R. Tolkien" is the creator of our example work;
4. Categories. They represent the genres or domains to which the artistic work belongs. *The lord of the Rings* categories may include "heroic fantasy", "Middle Earth universe" or "science fiction & fantasy";
5. Date of creation. *The fellowship of the ring (LOTR)* has been originally created in "1954".

The first three attributes are in most cases present in both work and entity. On the contrary, the last two attributes may lack in one or both data sources. Although the year of creation may be misleading, it is useful in specific cases. Dealing with the work about the movie *the lord of the rings : the return of the king*, there exist a first movie produced by Bass and Rankin in 1980 and a second one by Peter Jackson in 2003. If the creator's names are lacking or subject to mistakes, the dates could help us to disambiguate the two candidate movies. Finally, the idea is to compute the similarity for these five shared attributes of a work and an entity.

Computing Individual Similarity Values. We compute a similarity value between the same attributes of the work and the entity. However, the nature of

these attributes are different: the *title* and *creator* are plain text while the *categories* are a set of words. The *type* is a word from a finite set of values while the year can have different formats. As a consequence, we need different similarity measures for matching these attributes. Schema matching and ontology alignment research fields have provided many techniques to discover similar elements in various data sources that we can apply in our context [6].

Attributes Title and Creator. To measure the similarity between character strings, we have selected three terminological similarity measures: Jaro Winkler, Monge Elkan and Scaled Levenshtein. Combining these similarity measures enables us to avoid the drawbacks related to one of the measure (e.g., the Levenshtein returns high similarity for small-sized strings which are very dissimilar) [4]. Given the titles t (respectively creators c) of a work w and a LOD entity l, we compute the following similarity sim_{title} (resp. sim_{creat}) as the average between the three similarity measures:

$$sim_{title}(w,l) = \frac{jaro(t_w,t_l) + monge(t_w,t_l) + leven(t_w,t_l)}{3}$$

Attribute Categories. As these categories are represented by a set of strings, we define a very basic similarity function sim_{cat} between two sets. It computes the number of identical categories between the set of categories of a work w and a LOD entity l.

$$sim_{cat}(w,l) = \frac{|cat_w \cap cat_l|}{\max(|cat_w|,|cat_l|)}$$

Attribute Type. The type (extracted from its manifestations for the work) is limited to predefined values such as *book*, *movie*, *novel*. As the number of values is not large, we have built a small taxonomy extracted from the Wordnet hierarchy. To compute the similarity between two types, we can therefore apply the Resnik similarity [12]. It evaluates the similarity of these types based on the concepts that subsume them in our taxonomy. Figure 2 depicts a part of our taxonomy. For instance, the similarity value between the types *book* and *novel* in our taxonomy is equal to 0.29.

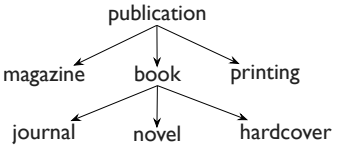


Fig. 2. A Fragment of our Taxonomy for Matching the Attributes *type*

Attribute Date. The idea is to extract only the year, which is a meaningful temporal granularity for artistic works. Thus, we compare the date value with several predefined patterns to extract the year, both for the work and for the

entity. If the extracted years from the work and the entity are identical, the similarity function returns 1. Else, it returns a 0 value. Back to our running example: Table 2 shows a LOD entity with each of its attribute's value. The last column indicates the similarity value for the attribute with regards to the corresponding attribute of the work (which is detailed in Section 4.1). We notice that the title and the creator are terminologically similar (similarity values around 0.8). As the work does not contain a date, the similarity value for creation date equals 0.

Table 2. Attributes and Similarity Values of the Entity *The_Fellowship_of_the_Ring*

Attribute	LOD Property Value	FRBR Work Attribute Value	Similarity Value
Title	The Fellowship of the Ring	The fellowship of the ring (LOTR)	0.77
Type	Book	Novel	0.29
Creator	J._R._R._Tolkien	JRR Tolkien	0.81
Categories	Fantasy	science fiction & fantasy	0.00
Date	1954-07-24	-	0.00

Computing a Global Similarity Value. From these attribute similarity values, we are able to derive a global similarity value. We have chosen a weighted average function to aggregate the values of all individual similarities. The global similarity value is computed with the following formula, where w is the work and l is the LOD entity, i.e., $w \in \mathcal{W}$ and $l \in \mathcal{L}$:

$$\text{sim}(w,l) = \frac{\alpha sim_{title}(w,l) + \beta sim_{type}(w,l) + \gamma sim_{creat}(w,l) + \delta sim_{cat}(w,l) + \zeta sim_{year}(w,l)}{\alpha + \beta + \gamma + \delta + \zeta}$$

In our running example, the DBpedia entity *The_Fellowship_of_the_Ring* and the work have a global similarity value equal to 0.37. As a comparison, the DBpedia entity related to the movie *The_Fellowship_of_the_Ring* obtains a similarity value of 0.22.

Filtering the Candidate Matches. Similarly to many matching approaches, we can filter the candidate matches by selecting those with a similarity value above a given threshold. A correct tuning of this threshold is crucial since it directly impacts the quality. Note that a constraint filter could also be applied in our context: if the work deals with a *movie*, then all LOD entities with a *book* type should be discarded. We demonstrate in Section 5 the impact of a threshold filter. As a result, all remaining entities discovered for a work can be ranked given their similarity values. Similarly to most matching approaches, the user still needs to decide if one of the proposed entities corresponds to the work. However, we show in our experiment results that our approach often ranks the correct entity at the first position.

4.4 Discussion

First, the LOD cloud is incomplete, i.e., it does not contain all entities that correspond to the FRBR works. Yet, our blocking process may return several

LOD entities, hence the need to compute their degree of similarity with the work. On the contrary, there may be no LOD entity returned by the blocking process. This does not mean that the LOD entity corresponding to the work does not exist. The benefits of our approach are threefold. First, it enables the verification of FRBRized data. But it can also be used to add new entities in the LOD cloud when a work has no corresponding entity. Specialized knowledge bases already use this mechanism to automatically create entities for a generic knowledge base, often with incomplete information. The last benefit is the semantic enrichment. Once a LOD entity is validated as correct for a given work, we can enrich this work by adding attributes extracted from the LOD entity. In addition, we can infer some simple relationships. For instance, we can link our work *The fellowship of the ring (LOTR)* with the other works of the trilogy thanks to the DBpedia property *dbpprop:books-of*. The attribute set we have chosen can be easily mapped to the attribute sets of different knowledge bases. If the number of attributes are too large for a manual mapping, tools such as Falcons [3] enable us to detect attributes that represent the same concept.

5 Experiments

In our experiments, a list of the 80 best selling fiction authors from Wikipedia[5] was used to query for product descriptions on Amazon bookstore (using the Amazon Product Advertising API[6]). These product descriptions have been FRBRized using the FRBRPedia approach [5,15], thus resulting in the generation of 684 distinct FRBR works. The challenge is to discover a correct entity on the LOD cloud for each of these works. In this experiment, we have chosen DBpedia as our main source of corresponding entities. Note that our approach is not limited to this knowledge base and that we could have used another source such as Freebase or OpenCyc. However, DBpedia is regarded as the center of this LOD cloud as it has the largest number of connections to other data sources.

5.1 Experimental Protocol

To reduce the search space, we could use SPARQL or Lookup queries. However, we have noticed that SPARQL queries are time-consuming with multiple constraints involving free-text strings. Thus, we use the *Lookup* API provided by DBpedia[7] to obtain a subset of DBpedia URIs representing entities that could correspond to the work using various queries as explained in Section 4.2. We used this reduced set of URIs as candidate matches for a given work. Matching techniques presented in Section 4.3 have been applied between the attributes of a work and those of the candidate matches. During this initial set of experiments, the global similarity value is computed with all weights equal to 1, which means that we do not promote any attribute. Similarly, we did not apply any filter to

[5] `http://j.mp/fiction_authors`, January 2011.
[6] `http://j.mp/amznProductAPI`, v.2010-10-01.
[7] `http://lookup.dbpedia.org`, December 2010.

this global similarity value (i.e., the threshold value is tuned to 0). The blocking and matching processes for the 684 works were performed in 10 to 12 minutes (without caching). Finally, we ranked the candidate matches for each work. For half of the 684 works, we were not able to discover any DBpedia entity. The remaining 343 works have at least one DBpedia entity. We presented the top-3 candidate matches for manual validation. This validation step was performed by 8 different people from our research group, which means that they have to check all proposed LOD entities and decide whether it corresponds to the given work (based on available information, such as creators, titles, summaries, or types). If none of the proposed entities is correct, participants validated the work by manually searching DBpedia. This manual validation forms a ground truth for the collection, based on which compute quality results of our approach.

5.2 Quality of Results

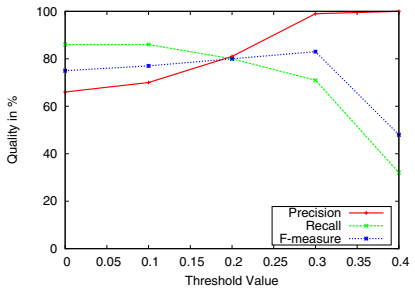
To assess the quality, we study the impact of the three parameters, namely the top-K matches, the threshold filter and the tuning of the weights in the global similarity value. Let us begin with the **top-K**. The number of correct discovered matches (true positives) at top-1, top-2 and top-3 are shown in Table 3. Most of the correct matches (189) are ranked at the top. At top-3, we only discover 12 more entities. Thus, our approach is able to present to the user the correct DBPedia entity at the top of the ranking.

Table 3. Number of True Positives by Top-k

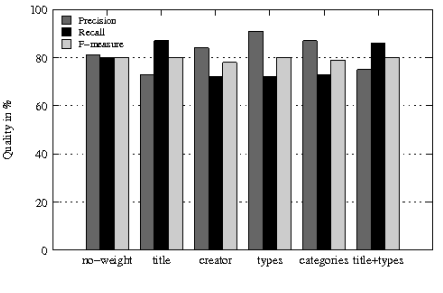
	Top-1	**Top-2**	**Top-3**
Number of True-Positives	**189**	197	201

The following experiment deals with **the impact of the threshold filter** (see Section 4.3). We compute the quality in terms of precision, recall and f-measure, as discussed in [6]. Precision represents the percentage of correct matches among those discovered at top-k while recall stands for the percentage of correct matches at discovered by our approach w.r.t. the total number of correct matches. F-measure is a tradeoff between precision and recall. Figure 3(a) depicts the quality obtained by our approach at top-1 when the threshold value for filtering matches varies. Without any threshold (value equal to 0), the f-measure reaches 76%. The recall value is around 85%, which means that we do not miss too many correct matches. However, we still discover many incorrect matches (precision at 66%). When we increase this threshold value, then the precision value increases while the recall score decreases. A balanced f-measure value (80%) is achieved for a 0.2 threshold. With higher threshold values, we are able to reach 100% precision, but at the expense of recall (71%). A peak is reached when the threshold is in the range of 0.3 and 0.4.

In the last experiment we study **the impact of weights in the global similarity function** (see Section 4.3). We have previously shown that a threshold value equal to 0.2 provides balanced results between precision and recall, so we



(a) Quality Results (precision, recall, f-measure) w.r.t. a Threshold Filter



(b) Quality Results w.r.t. the Weights of Individual Similarities

Fig. 3. Quality Results

have used this value in this experiment. Figure 3(b) depicts the top-1 quality when we apply a higher weight to one or more individual similarity measures. For instance, the precision, recall and f-measure values respectively equal 73%, 87% and 80% with a weight on the *title* similarity measure. We notice two interesting points. The former deals with a weight on the title which enables the promotion of recall (87%). Indeed, when a work matches a LOD entity, their titles are often similar. But this high title similarity is limited by the other individual similarity functions. Thus, tuning the weight of the title allows us to discover more correct matches, but at the expense of precision. The latter point is the weight applied to types which promotes precision (91%). Indeed, a hard constraint on the types avoids the discovery of matches involving a work and a LOD entity with different types (such as movie and book).

5.3 Discussion

Our first observation is concerned with the quality of the input data and of their conversion into FRBR. This process obviously has an impact when linking to DBpedia. For instance, the search results from Amazon need to be cleaned. Indeed, they can contain dirty data such as "The Lord of the Rings: The Return of the King (Widescreen Edition)" and unrelated products (given the query). We also faced several issues with the DBpedia knowledge base. A lack of information in the DBpedia entity leads to no match, a case which may occur for DBpedia entities which are automatically created from other knowledge bases but with incomplete attributes. Similarly, an entity page can redirect to a related entity page (e.g. author, concept, event). As for the experiment results, our global similarity measure is reliable since most correct matches are discovered at top-1 and we miss only a few entities. Furthermore, the approach is flexible with the weights/threshold, which both enable users to promote either precision or recall.

6 Conclusion

In this paper, we have presented a generic framework to link a FRBR work to its corresponding LOD entity, using a query builder as blocking process and

refined similarity measures as matching process. As a result of experiments with Amazon, we have successfully discovered the correct DBpedia entity for most products. Thus, our approach is a basis both for verification purposes and for semantic enrichment. As for future work, the framework can be integrated with other LOD data sources (e.g. *Freebase*, *LastFM*). Indeed, linking a work to a specialized database (e.g. *MusicBrainz* for musical work) may provide a higher probability for discovering the correct match than a general knowledge base.

References

1. Aalberg, T., Haugen, F.B., 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. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (2009)
3. Cheng, G., Ge, W., Qu, Y.: Falcons: searching and browsing entities on the semantic web. In: Proc. of WWW, pp. 1101–1102 (2008)
4. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: Proc. of IIWeb (2003)
5. Duchateau, F., Takhirov, N., Aalberg, T.: FRBRPedia: a Tool for FRBRizing Web products and Linking FRBR Entities to DBpedia. In: Proc. of JCDL (2011)
6. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
7. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association 64, 1183–1210 (1969)
8. Freire, N., Borbinha, J.L., 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., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 267–276. Springer, Heidelberg (2007)
9. Geva, S., Kamps, J., Trotman, A.: Focused retrieval and evaluation, INEX 2009, vol. 6203. Springer, Heidelberg (2010)
10. IFLA Study Group on the FRBR. Functional requirements for bibliographic records, final report. UBCIM Publications; New Series, 19(1) (1998)
11. Köpcke, H., Rahm, E.: Frameworks for entity matching: A comparison. Data Knowl. Eng. 69, 197–210 (2010)
12. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
13. Riley, J.: Enhancing interoperability of frbr-based metadata. In: Dublin Core and Metadata Applications (2010)
14. Rode, H., Serdyukov, P., Hiemstra, D.: Combining document- and paragraph-based entity ranking. In: SIGIR, pp. 851–852 (2008)
15. Takhirov, N., Duchateau, F., Aalberg, T.: Supporting frbrization of web product descriptions. In: Proc. of TPDL, Springer, Heidelberg (2011)
16. Vercoustre, A.-M., Thom, J.A., Pehcevski, J.: Entity ranking in wikipedia. In: SAC, pp. 1101–1106 (2008)
17. Voorhees, E., Harman, D.: Trec experiment and evaluation in information retrieval. MIT Press, USA (2005)

Interactive Vocabulary Alignment

Jacco van Ossenbruggen[1,2], Michiel Hildebrand[1], and Victor de Boer[1]

[1] VU University Amsterdam
[2] CWI Amsterdam, The Netherlands

Abstract. In many heritage institutes, objects are routinely described using terms from predefined vocabularies. When object collections need to be merged or linked, the question arises how those vocabularies relate. In practice it often unclear for data providers how well alignment tools will perform on their specific vocabularies. This creates a bottleneck to align vocabularies, as data providers want to have tight control over the quality of their data. We will discuss the key limitations of current tools in more detail and propose an alternative approach. We will show how this approach has been used in two alignment use cases, and demonstrate how it is currently supported by our Amalgame alignment platform.

1 Introduction

In the library, archive, museum and many other domains, objects are routinely described using terms from predefined vocabularies. When object collections need to be merged or linked, a typical question that needs to be answered is how those vocabularies relate. More specifically, one would like to know which concepts from different vocabularies correspond to one another. We will call a set of such correspondences an *alignment*.

There is an active research field that studies methods and techniques to generate alignments automatically. We experienced that in practice it is, however, difficult to apply these techniques to vocabularies in the cultural heritage domain. Most alignment tools are not designed for the large but shallow vocabularies typical in this domain. Furthermore, tools provide little support to analyse large sets of correspondences, making it difficult to assess the quality of the generated results. To tackle these issues we propose an interactive approach to vocabulary alignment.

In the next section we discuss the limitations of fully automatic alignment tools. In Sect. 3 we describe a semi-automatic, interactive approach. In Sect. 4 we will show how this approach has been used in two alignment use cases, and demonstrate how it is currently supported by our *Amalgame* alignment platform[1]. Finally, we reflect on our approach and discuss future work.

[1] Amalgame is open source and available at `http://semanticweb.cs.vu.nl/amalgame/`. All alignment strategies discussed in the use cases have been published at `http://semanticweb.cs.vu.nl/lod/tpdl2011/` and can be "replayed" in Amalgame, allowing full replication of all alignments described in this paper.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 296–307, 2011.

2 Problem Analysis

There is an active research field that studies methods and techniques to generate alignments automatically, and the tools produced by this field are evaluated yearly in the context of the Ontology Alignment Evaluation Initiative (OAEI)[2]. A key insight from this field is that two concepts can be similar or dissimilar along many different dimensions [3]. Automatically finding similar concepts typically requires some hybrid approach that combines different techniques, each addressing a part of the total set of potentially interesting dimensions. Another important insight is that the application context in which the alignment will be deployed often influences what constitutes a "good" alignment [5]: two concepts might be regarded as sufficiently similar in one context, but not in another. The main approach in vocabulary alignment is to develop hybrid tools that try to fully automatically find some smart combination of techniques to generate an alignment, and (b) allow the developer to tune the tool so that the alignment fits a specific application context.

While the approach sketched above is well established, both results from our previous work [9,7,8] and feedback received from domain experts during our work in the MultimediaN E-Culture[3], Europeana(Connect)[4] and PrestoPrime projects[5], indicate that it also has some major limitations when it has to be applied in the cultural heritage field.

First, domain experts find it hard to determine how well a tool would perform for their alignment task. From the alignment research literature, it is clear how each tool performs on the data used in the evaluation experiments. However, due to the complexity of the good performing tools, it often remains unclear *why* some tools perform better than others, so it is hard for experts to predict which tool would be suitable for their own data set.

Second, experts perceive the current tools to not support the large and shallow vocabularies that are typical for their domain. Most alignment tools target complex vocabularies with different ontological relations, but only several 100s or 1000s of classes. In the cultural heritage domain the vocabularies typically contain only a few thesaurus relations, but frequently contain over 10,000s or even 100,000s of concepts. When run on larger vocabularies, many tools simply crash, or fail to finish alignment runs within a reasonable amount of time.

Third, when a tool finishes successfully, it typically produces a result set with a large number (e.g. over 100k) of correspondences, but provide little support to assess the quality of these results. Furthermore, the quality of the correspondences might not be homogeneously distributed across the alignment result set. Different subsets of alignments might have different features that determine the quality of the end result. Transparent and interactive assessment is crucial to be able to decide whether the result is of sufficient quality.

[2] http://oaei.ontologymatching.org/
[3] http://e-culture.multimedian.nl/
[4] http://www.europeanaconnect.eu/
[5] http://www.prestoprime.org/

Fourth, when the results are not sufficient it is unclear how the tool should be (re-)configured to improve the results. Experts need to be able to understand why a tool found erroneous correspondences and how to get rid of them in a next step to improve precision. When the tool failed to find correct correspondences, the experts need to know how to find those in a next step to improve recall. This often requires insight in how the alignment algorithms work, and how to configure them to adjust them to the specific needs of vocabularies at hand.

Remark that the first two problems are related to the fact that fully automatic alignment tends to result in complex techniques that are relatively slow on large data sets and hard to explain to domain users. The last two problems are due to the fact that current tools are designed to produce an alignment of sufficiently high quality in a single run, without much input from the user, while in practice experts feel that the required quality can only be achieved by multiple runs, where each run requires their input.

In the next section, we sketch an alignment approach that is based on these insights. We then show the feasibility of our approach by discussing two use cases of vocabulary alignments in which we have used this approach.

3 The Amalgame Approach to Vocabulary Alignment

To address the problems above, we developed an alignment approach that improves the speed and transparency of the alignment process by drastically reducing the complexity of the technology, allowing the user to combine a limited number of basic building blocks into an alignment workflow targeted to the data set at hand. Each building block should be sufficiently simple to produce an understandable result. Which blocks to use and in what order or combination is fully controlled by the user. Furthermore, produced alignments (both intermediate and end results) can be easily evaluated to give insight in their quality.

We have built a prototype alignment service that has been designed with this approach in mind, and used the prototype to create alignments in two different use cases, that will be discussed in the next section. Here we sketch an high level overview of the Amalgame alignment methodology and will flesh out some interesting details in the context of the use case descriptions.

3.1 Vocabulary Analysis

An assumption of the interactive approach is that the user has knowledge of the vocabularies being aligned. Here, we focus on vocabularies that can be represented by SKOS [6]. For such SKOS-like vocabularies we identify two types of characteristics. First, the user needs to know how the vocabularies differ in size and heterogeneity. Second, the user has to identify the concepts' properties that can be used in string matching. Third, the user has to identify other properties that can be matched, such as hierarchical and associative relations.

3.2 Workflow Components

Our approach is to have the user interactively construct an alignment workflow. The individual building blocks of this workflow consist of: **selectors** to define which concepts to use from the source and target vocabularies, **matchers** to find correspondences between the selected source and target concepts, **partitioners** to split sets of correspondences, **mergers** to create unions of specific subsets, **analyse tools** to investigate the mappings, and **filters** to select specific correspondences and discard others.

3.3 Interactive Alignment

Alignment within Amalgame is a process where the user iteratively applies matchers, partitions the result set, and applies new matchers or a filter. After each step the user typically analyzes the results to determine the next step. We identify five typical scenarios, depending on the outcome of the analysis.

- The first scenario is that a user decides the results are no good at all, in which case all results are simply discarded after analysis. Assuming the technique used is sufficiently simple, the user will understand from the analysis what caused the failure and will be able to try another matching run, using another technique or a better configuration of the technique used in the previous run.
- The second scenario is that the results are good, but that recall is low. To improve recall, the user can proceed by matching only the concepts that have not yet been aligned. Note that this result set is typically a smaller set, so the user may decide to deploy computationally more expensive matching techniques to improve recall in subsequent runs.
- The third scenario is that the results are good, but that precision is low. To improve precision, users need to find filters that allow them to distinguish true from false correspondences. Again, more expensive techniques can be used to boost precision for smaller subsets.
- The fourth scenario is that a user decides that the results are of sufficient quality, after which she exports them to the desired format and we consider the alignment task to be successfully finished.
- The fifth scenario is that the user finds the results of insufficient quality, but is out of options and does not know how they can be further improved, in which case we consider the alignment task to be failed.

In practice, we found the first scenario useful to quickly try some alternative matchers, and to compare, analyse and discard the results, just to develop some intuition before the real alignment task starts. Many alignment tasks, including the first two use cases discussed below, are based on an iteration of the second and third scenario. Ideally, with each iteration the set of concepts that have to still be aligned (to improve recall) and the set of correspondences that still have to be filtered (to improve precision) decreases, or, if not, the user gains some knowledge to achieve this in the next step.

4 Use Cases

In this section we describe two alignment use cases. We found that, in practice, the in-house vocabularies from different institutes are sometimes directly aligned with each other, but typically they are indirectly related by aligning them to the same external vocabulary. As the first use case we explore such an alignment of an in-house vocabulary to an external vocabulary. We consider the alignment of the thesaurus of the Netherlands Institute for Sound and Vision, GTAA, with a general linguistic vocabulary of Dutch, Cornetto. A benefit of an alignment with such an external vocabulary is that this also makes the alignments of this vocabulary available for the in-house vocabulary. For example, Cornetto already contains links to the English WordNet. A different example where alignment is required, is when a new version of a vocabulary is released, and no direct links between the two are maintained. In the second use case we consider the mapping of two different versions of WordNet. The two use cases show typical examples of one-to-one, one-to-many and many-to-many correspondences (abbreviated as *1-1*, *1-n* and *n-m* below). While our approach could be applied to a wide variety of mapping relations, the use cases focus on relatively simple, bi-directional equivalence relations. More complex relationships, e.g. as described in [2], could be addressed by either deploying more complex workflows or more manual intellectual input.

4.1 In-House to General: GTAA to Cornetto

The Netherlands Institute for Sound and Vision uses an in-house thesaurus for the documentation of audiovisual content. This so-called GTAA thesaurus (Dutch acronym for Common Thesaurus Audiovisual Archives) contains approximately 160,000 terms in six facets: subjects, locations, person names, organization names, maker names and genres. In this use case we focus on the terms in the subjects facet.

Cornetto is a WordNet-like lexical semantic database of Dutch that contains 70,000 synsets [10]. Compared to the GTAA subject terms, the synsets provide a large number of additional synonyms and an extended description. The synsets are linked into an elaborate hierarchical structure.

To goal of making the alignment is to improve Dutch access to the institute's collection by taking advantage of Cornetto's additional labels (e.g. synonyms) and semantic relations to GTAA's subject terms. In addition, the existing alignment between Cornetto and WordNet could also provide an English access point to the archive.

For this use case we map the GTAA subject terms to Cornetto synsets. As Cornetto contains the same words in different synsets (e.g. homonyms), we can expect that string matching techniques will find multiple synsets for many GTAA subject terms. Our focus is to choose the right target synset(s) for each source. Typically, this will be one synset per GTAA subject term (that is, *n-1* correspondences), but there might be cases where multiple synsets are good candidates. In this case, the aim is to find not the best, but all correct targets (that is, *n-m* correspondences).

Vocabulary Analysis. We start the alignment process with an exploration of the GTAA subject terms. In total there are 3,932 subject terms. All terms have at least one preferred label, often an alternative label and one or more related terms, and some have a description. In addition, the subjects are organized in an hierarchical structure. We observe that the majority of the terms are nouns. In Cornetto this part of speech distinction is explicit, as each synset is of word type: noun (52,845), verb (9,017), adjective or adverb. Ideally, we would like to map the nouns in GTAA to the nouns in Cornetto. However, there is no explicit information in GTAA to automatically distinguish the nouns from the verbs. We choose the next best solution and start with the alignment of all GTAA subject terms to the nouns in Cornetto. We assume that there will be no or very few verbs from GTAA that will be incorrectly mapped to the nouns in Cornetto.

We also observe that the most labels of the GTAA subject terms are in plural form, whereas the labels in Cornetto are in singular form. When matching the labels we should account for this difference. Finally, we observe that where GTAA discriminates between preferred and alternative labels, Cornetto only has one type of label, which has been mapped to `skos:altLabel`.

Interactive Alignment. Given the discussions above, it is not *a priori* clear which string matching strategy to use. We expect that using alternative labels, in addition to the preferred labels, will increase recall, but are unsure at what expense (in terms of precision). Similarly, we expect that stemming will deal with the plural GTAA nouns and singular Cornetto nouns, but it might also introduce new problems. We decide to explore different options and try matching including and excluding GTAA alternative labels. We also match with and without stemming.

Table 1. Number of correspondences between GTAA and Cornetto. Horizontally, the labels used: preferred labels only and including alternative labels. Vertically, the label similarity metric: exact matching or matching after stemming.

	Preferred labels			Preferred + alternative labels		
	total	n-1	n-m	total	n-1	n-m
exact	1,190 (30%)	880	310	1,319 (33%)	829	490
stem	2,493 (63%)	1785	708	2,725 (69%)	1655	1070

Table 1 shows the statistics for the different string matching techniques[6]. From the column labeled *total*, we observe that there is indeed a large increase when stemming is used. We can also observe that by including the alternative labels more correspondences are found. Based on these observation we might opt for the approach that gives us the highest recall: matching the stems of both the preferred and alternative labels. Before we make this decision there is, however, another important characteristic of the results that we should consider. How

[6] The mappings generated in this use case can be found online at `http://semanticweb.cs.vu.nl/lod/tpdl2011/gtaa_cornetto`

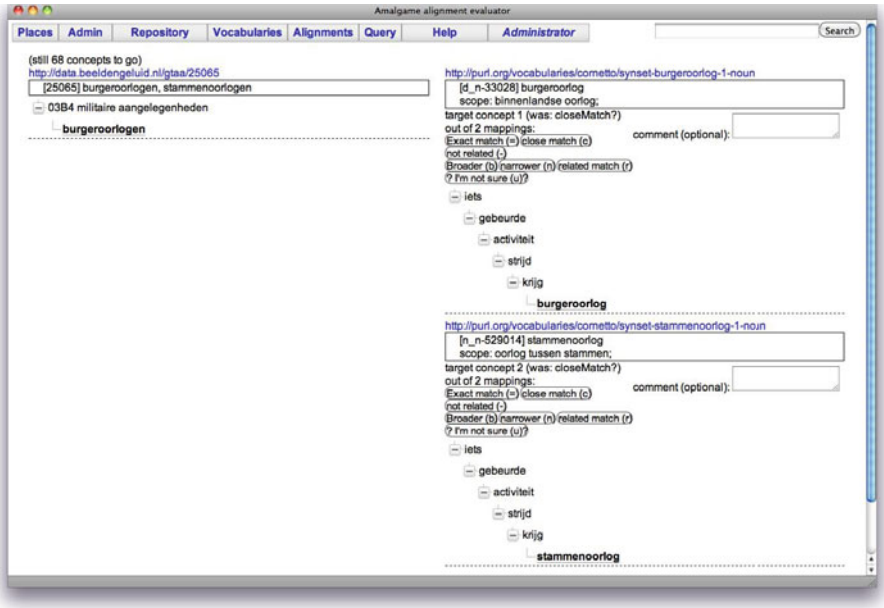


Fig. 1. Screenshot of the Amalgame evaluation prototype. On the left the source concept from GTAA and on the right two target concepts from Cornetto. The GTAA concept "burgeroorlog" (dutch for civil war) with alternative label "stammenoorlog" (dutch from tribal war) is mapped to two different targets. The target concepts in Cornetto, civil war and tribal war, are siblings as they are two specific types of war.

many target concepts are found for each source concept? And in case multiple targets are found, is this caused by ambiguity of the source concept or are all targets valid alternatives?

To investigate different types of alignments we use Amalgame to partition the set of correspondences. We partition them in a set where the source concepts have only 1 target, and another set where the source concepts have multiple targets. Table 1 lists the number of sources that are mapped to only 1 target (the *n-1* column includes both $1-1$ and true $n-1$ results, and mapped to multiple targets (idem, *n-m* also includes $1-m$). We observe that the number of *n-1* alignments is larger when only the preferred labels are included. In other words, the alternative labels primarily introduce extra targets for sources that were already mapped. Do these alignments introduce unnecessary ambiguity, or are the additional targets valid alternatives?

To analyze the results in more in detail we use Amalgame to visualize correspondences including the relevant information of the source and target concepts. In this case, we are interested in the *n-m* mappings introduced by the alternative labels. We produce this set by subtracting the 708 *n-m* correspondences found by matching preferred labels from the larger set of 1070 *n-m* found by matching both preferred and alternative labels. From the resulting set we take a random

sample of 25 correspondences to investigate in detail. Figure 1 shows a screen shot of this investigation. For a single source concept it lists the multiple target concepts. In addition, all alternative labels, descriptions and related terms are shown. Going through the sample set we found four different types of *n-m* correspondences:

1. One of the targets is more generic then the others. Cornetto is more fine-grained then GTAA. A single concept in GTAA containing multiple labels, e.g. "poison, pesticide", is mapped to different targets in Cornetto, where "poison" is more generic than "pesticide". In this case we want to select the most generic term. Optionally, we could create narrower matches between the other targets, but this is outside the scope of this paper.
2. The targets are siblings of each other. Again the granularity difference between the vocabularies often causes a single concept in GTAA, e.g. "civil war and tribal war", to be mapped to different targets in Cornetto, where "civil war" and "tribal war" are siblings as they are more specific types of "war" (shown in Figure 1). In this case all targets are valid alternatives and we want to keep an 1-n correspondence to all siblings.
3. The targets are about the same topic. Some concepts in GTAA contain labels for different types of things, but related in topic e.g. "beekeeping and honey combs". In Cornetto these are different terms in completely different parts of the hierarchy. We choose again to keep all targets and create a 1-n correspondence. If we would have the rights to modify GTAA, we could also decide to split the source concepts into two separate concepts.
4. The targets are different senses of the source concept. A GTAA concept is matched to one concept from Cornetto by its preferred label and to another by its alternative label. For example, by the preferred label "capitulate" a single concept from Cornetto is found. By the alternative label "surrender" it finds the same concept, but also the concept that refers to "surrender of attention". In this case the source concept is ambiguous and only one target should be selected.

We conclude that by using only the preferred labels valid alternatives are excluded. Therefore, we choose to include alternative labels and match them after stemming. The *n-1* correspondences generated with this configuration are likely to be correct, as we used a simple matching algorithm that fits well with the labels in our vocabularies. Evaluation of a random sample of 25 confirms this assumption, as all correspondences are indeed correct. At the other hand we have a larger set of *n-m* correspondences. The analysis of this set provided us with a number of different cases. How can we use this knowledge to find the valid *n-m* correspondences and, in case of ambiguity, select the best candidate to get the *n-1* correspondence we are looking for?

To automatically detect different types of correspondences and select the best target candidates Amalgame provides a number of strategies. We configure these strategies for the different types of *n-m* correspondences. We start with the *n-m* set (1070 source concepts) and try to identify the correspondences for each case.

For 91 source concepts we can find a target that is more generic than the other targets. These concepts are found by configuring the Amalgame partitioning component to check for hierarchical relations between the targets. From the remaining correspondences, 72 sources have sibling targets.

For the remaining *n-m* correspondences, we try to automatically detect the most suited candidate. We observed that the wrong targets can occur in different sub-trees of Cornetto. Therefore, we can identify the best target by the hierarchical similarity to the source target. For each ambiguous correspondence we check if the source and target have similar ancestors or descendants. To test for similarity between the terms in the hierarchy we use as a base set the *n-1* correspondences. When the hierarchy of one target has more aligned concepts with the hierarchy of the source it is a better candidate. As this method adds new correspondences, it extends the base set, possibly relevant for further disambiguation. Therefore, we repeat this procedure until no more additional matches are found. In total, for 342 source concepts we manage to find a distinguishing target.

Finally, we decide to align all remaining GTAA subjects to the verbs in Cornetto. Analysis of the vocabularies also makes clear that the labels of the verbs, in both vocabularies, are in infinitive form. Therefore, we choose to align them using exact string matching. For 115 source concepts we find correspondences, 78 of these are *n-1* mappings, while 37 are *n-m* mappings. As the *n-m* set is very small, we can manually evaluate it. Within 14 minutes we manually disambiguated 19 sources, and accepted multiple alternatives for two sources. For the remaining 14 source concepts we decided they were falsely mapped. All were nouns that were not mapped due to limitations of the stemming algorithm. We expect the same stemming problem causes errors in the set of *n-1* correspondences, and also manually evaluate these. Within only 5 minutes we found the 13 source concepts were it went wrong.

Results. In total we found matches for 2275 (58%) concepts from the GTAA subjects facet. From these the large majority (2160, 55%) were matched to Cornetto nouns. For 42% of the GTAA subjects we found a correspondence to only one target. As we used a simple matching technique, we expected high precision for this subset. In an evaluation of a small sample of this set all correspondences were judged to be correct. In the remaining set we identified four ways in which multiple targets were found. We configured the filter components to identify these cases. For more than half of the 1-n matches we managed to either select the best target or confirm that all targets are valid alternatives. To judge the other half of the matches manual evaluation is required. In future work we would like to perform such an evaluation with the users of GTAA. Finally, only 115 GTAA subject terms were mapped to Cornetto verbs. This small set we manually evaluated in only a few minutes.

4.2 Versioning: WordNet 3.0 to WordNet 2.0

WordNet is a large lexical database of English published by Princeton University. It groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms

(synsets), each expressing a distinct concept [4]. W3C released an RDF version of Princeton's WordNet 2.0 in June 2006 [1]. In August 2010 we released an RDF conversion of Princeton's WordNet 3.0 as Linked Open Data. Until now, there is no reliable data set that specifies which synset in version 3.0 correspond to which synset in version 2.0. A typical goal of creating these correspondences would be to update corpora indexed with the old version to the new version.

Vocabulary Analysis. To be able to treat WordNet as a SKOS vocabulary, we use a simple schema mapping: WordNet synsets are mapped to SKOS concepts, WordNet sense labels to SKOS altLabels and WordNet glosses to SKOS definitions. WordNet 2.0 consists of 115,424 concepts with a total of 203147 labels. WordNet 3.0 has slightly more (117,657) concepts with 206,976 labels.

Because WordNet maintenance is largely a manual effort, we expect many concepts will have remained the same and will be easy to map. Concepts that we will choose to leave unmapped are those 2.0 synsets that have been dropped in the new version, without having a counterpart in the new version and the 3.0 synsets that are newly added without having a counterpart in the old version. Concepts that we would like to map but could be hard to do automatically include concepts that have splitted or merged between versions, and concepts of which so many properties have changed that it is hard to tell if we are dealing with the "same" concepts or not.

Both vocabularies are splitted into nouns (70%), verbs (12%), adjectives (15%) and adverbs (3%). We assume that by mapping only nouns to nouns, verbs to verbs, etc. we can both reduce the search space and avoid many erroneous mappings between homonyms in different parts of speech. This approach risks missing concepts that moved to another part of speech category, but we assume this to occur very infrequently or not at all.

Interactive Alignment. When aligning WordNet 3.0 to 2.0 we would like to explicitly use our knowledge of the fact that we are aligning two versions of the same vocabulary. For example, given the large amount of homonymy, we expect a simple label match to produce many correspondences, most of which will be wrong. In contrast, we expect the definitions to be unique for most concepts, and since manually updating many definitions is hard manual work, we expect the majority of the concepts to have the same definition in both versions.

So as a first step, we try a quick case insensitive match on skos:definition. Selecting only the *1-1* mappings results leaves us with 103,521 correspondences (set 1a[7]), covering already 89.7% of all 2.0 synsets. Of the *n-m* correspondences, 931 can be reduced to *1-1* (set 1b) by simply matching also the labels. We quickly evaluate the remaining 26 correspondences (set 1c) manually, and conclude these are all cases with duplicate synsets in one or both versions, so all the 26 remaining *n-m* correspondences turn out to be correct too. After this simple first step, we only need to align less than 10% of the original number of concepts, so we can afford more expensive techniques in the following steps.

[7] See online results at `http://semanticweb.cs.vu.nl/lod/tpdl2011/wn3020`

As a second step, we run a cheap, case insensitive label match on the remaining concepts. This yields another 6379 *1-1* correspondences (set 2a), which we assume to be mostly correct. As expected, it also results in a relatively large number of *n-m* correspondences: 8528 matches between only 3502 source and 3319 targets. In this set we thus expect many wrong homonym matches. We run a more expensive string distance matcher on the definitions, after which we select, for each 2.0 target concept, the source with the most similar definition. This reduces the set to 3310 mappings between 2807 sources and 3310 targets, for only 9 targets we find 19 mappings to two or more equally similar sources. A quick manual evaluation found that only 8 of these 19 were correct (set 2b). Repeating this step in the other direction, be selecting for each of the 2807 source the most similar target, we find 2800 *1-1* mappings (set 2c), with 14 mappings for the 7 sources for which there two equally similar targets. Manual evaluation found 9 of these correct (set 2d).

Results. We have created three distinct subsets of correspondences in the first step and four subsets in the second step. Together, these seven sets consist of 113,675 correspondences for a similar number of WordNet 2.0 concepts, covering 98.48% of all 2.0 synsets. For each subset, we can easily describe how it has been created, and why we would or would not trust the correspondences they contain. A more thorough manual evaluation could take this into account, by taking strategic samples from each subset. The coverage can be further increased by trying to map concepts for which (all) the labels have been changed between versions, as happens when spelling errors are detected or new spelling conventions are applied, but this is out of scope for this paper.

5 Discussion

We conclude it is feasible to construct an alignment workflow for relatively large SKOS-like vocabularies by combining simple techniques. With the prior knowledge of the vocabularies and analysis of the correspondences we iteratively increased recall and precision. The resulting alignments are comprised of multiple homogeneous subsets of correspondences. This allows for targeted evaluation per subset. In addition, this allows to combine evidence from multiple subsets to increase precision, or strategically select multiple subsets to increase recall.

A potential drawback of our approach is that the selection, configuration and combination of components is the responsibility of the user. This makes the approach less attractive for data sets were fully automatic approaches produce results of sufficient quality. A potential risk is that we assume a finite and relatively small set of basic components. Amalgame currently provides a number of such components, some of these were used across use cases. During the specific use cases, however, we also found a need for additional components. Creating these components was straightforward. New use cases might require new components as well.

The workflows for the use case presented in this paper were created by the authors, using an experimental interface. Our longer term goal is to support

vocabulary owners to create their own alignments. This requires a user interface to iteratively construct alignment workflows. Currently we are developing such a user interface. The interface combines the construction of a workflow, with the analysis of mappings. Thus, each time extending a single node and using the analysis tools to investigate intermediate results. In future work we will evaluate such an interface with the vocabulary owners.

Acknowledgements. We thank W. van Hage, A. Isaac, C. Reverté Reverté, A. Tordai and J. Wielemaker for their feedback and help in the development of Amalgame. M. van Assem produced the RDF conversions for WordNet 2.0 and 3.0. This work was partially supported by the PrestoPRIME and Europeana-Connect project.

References

1. van Assem, M., Gangemi, A., Schreiber, G.: Conversion of WordNet to a standard RDF/OWL representation. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy (May 2006), http://www.cs.vu.nl/~mark/papers/Assem06a.pdf
2. Doerr, M.: Semantic problems of thesaurus mapping. J. Digit. Inf. 1(8) (2001)
3. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg, (DE) (2007)
4. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series. MIT Press, Cambridge (1998)
5. van Hage, W.R., Isaac, A., Aleksovski, Z.: Sample evaluation of ontology-matching systems. In: Garcia-Castro, R., Vrandecic, D., Gómez-Pérez, A., Sure, Y., Huang, Z. (eds.) EON. CEUR Workshop Proceedings, vol. 329, pp. 41–50. CEUR-WS.org (2007)
6. Miles, A., Bechhofer, S.: Skos simple knowledge organization system reference. W3C Recommendation (August 18 (2009), http://www.w3.org/TR/skos-reference/
7. Tordai, A., van Ossenbruggen, J.R., Ghazvinian, A., Musen, M.A., Noy, N.F.: Lost In Translation? Empirical Analysis Of Mapping Compositions For Large Ontologies. In: Proceedings of 5th International Workshop on Ontology Matching 2010. CEUR-WS (November 2010)
8. Tordai, A., van Ossenbruggen, J.R., Schreiber, G., Wielinga, B.: Aligning Large SKOS-Like Vocabularies. In: ESWC 2010. LNCS, vol. (7), pp. 198–212. Springer, Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-13486-9_14
9. Tordai, A., van Ossenbruggen, J., Schreiber, G.: Combining vocabulary alignment techniques. In: K-CAP 2009: Proceedings of the Fifth International Conference on Knowledge Capture, pp. 25–32. ACM, New York (2009)
10. Vossen, P., Maks, I., Segers, R., van der Vliet, H.: Integrating lexical units, synsets and ontology in the Cornetto database. In: (ELRA), E.L.R.A. (ed.) Proceedings of the Sixth International Language Resources and Evaluation, LREC 2008 (2008)

The Impact of Distraction in Natural Environments on User Experience Research

Elke Greifeneder

Berlin School of Library and Information Science, Humboldt-Universität zu Berlin,
Unter den Linden 6, 10099 Berlin, Germany
greifeneder@ibi.hu-berlin.de

Abstract. Laboratories have long been seen as reasonable proxies for user experience research. Yet, this assumption may have become unreliable. The trend toward multiple activities in the users' natural environment, where people simultaneously use a digital library, join a chat or read an incoming Facebook post, changes users' behavior. The effects of these disruptions generate a gap that is generally not taken into account in user-experience research. This paper presents a psychological experiment that measured how differently people behave in a laboratory and in a natural environment setting. The existence and impact of distraction is measured in a standard laboratory setting and in a remote setting that explicitly allows users to work in their own natural environment. The data indicates that there are significant differences between results from the laboratory and natural environment setting. Distractions like email or chat influence the users' performance and their ratings.

Keywords: User studies, digital library, distraction, laboratory, remote evaluation, natural environment.

1 Introduction

Laboratories have long been seen as reasonable proxies for user experience research. Yet, this assumption may have become unreliable. The changes in the digital world have resulted in many shifts in user's behavior. Mobile technology offers any-time any place worldwide access to information services. Users can access services in a library, at work, at home, as well as in cafés, in the metro or at noisy festivities. Their natural use environment differs at least in one major way from laboratories: distractions are ubiquitous.

In laboratories, distraction is eliminated as a confounding variable and not treated as an existing influential part of the users' information environment. The trend toward multiple activities in the users' natural environment (some researchers call it also real-life or real-world environment), where people simultaneously use a digital library, join a chat or read an incoming Facebook post, changes users' behavior. The effects of these disruptions generate a gap that is generally not taken into account in user-experience research.

The current paper seeks to close this gap and to examine whether proactive applications – like social networks, email clients or mobile phones – and other disturbing factors distract users and change their behavior in every day interactions with digital services.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 308–315, 2011.

2 Research Background

Laboratories have many advantages. Because of the absence of confounding variables, laboratories allow researchers to work in a controlled environment and to assign a particular phenomenon to a single concrete behavior. But worldwide access to information services has led to a greater variety of users – with a broad spread of cultural backgrounds. Bringing a representative sample group to the laboratory could be a costly challenge based on travel expenses alone.

In response to the limitations of laboratories, researchers started to discuss the impact of the information context. While there are many studies on the information use environment (for an older but much cited review, see [7]), it is a broader concept than the natural environment and research about it has not explicitly studied the impact of distraction on users' behavior.

Remote studies are another way of dealing with the limitations of laboratories. In a very broad sense, remote implies a distance between the participant and the researcher. In a synchronous remote setting (also called a moderated remote test), researcher and participant are separated in space, but they have a real-time connection using text, voice or video. This mode of remote study was heavily used in early forms of remote usability testing. Its advantages over laboratory tests are the elimination of travel costs and the familiar environment for participants. Asynchronous tests (also called unmoderated tests) came up more recently and add a temporal dimension. Researchers and participants are now separated in time as well as place. Participants have no direct contact with the researcher and can access the test at their place and time of convenience. With asynchronous tests, bigger samples are also possible. [3] offer more details on the two methods and describe earlier studies.

Remote tests are widely used in information retrieval and in usability research. About 30 studies have been conducted to examine whether one can replace laboratories with remote settings (see [3] or [6]). Both fields report that there is no significant difference between the two settings. Only few studies like [9] report a difference in the number of usability issues. It is important to note that [9] also discovered that participants tend to give more positive ratings in the laboratory than in the remote setting.

The current study describes a remote test that has specifically been adapted to embrace the users' natural environment. Remote does not necessarily equal natural environment. Many studies design their remote tests to be virtual laboratories because minimizing potential external factors makes the statistical analysis more straightforward. [9] deleted the data of participants that took over 1000 seconds, which they saw as an indication of an interruption. In a second experiment, they even integrated a pause button for participants, but they gave no information about how frequently this button was hit. [5] asked participants to close all other applications on their computers and not to talk to others. Apart from a study about the impact of stress ([2]), remote performance tests rarely observe softer elements like distraction.

Distraction has a real influence on behavior. The time to complete a test is a factor that is used in psychology to measure the level of distraction. [4] studied the effect of distraction on completion time rates. They observed two groups in a laboratory: one group was disturbed twice by an incoming chat. The participants had to search in a book list with two levels of difficulty: find specific book titles and find books on specific themes. They collected data on completion time; in the chat group, they removed the chatting time.

This experiment made four important discoveries: 1) distracted participants take more time to complete a task; 2) there is a statistically significant difference between the two groups; 3) the difference in time is not entirely related to changes between keyboard and mouse, but comes from the influence of distraction on memory, which relates to task difficulty; 4) distractions "reliably harm faster, stimulus-driven search tasks more than effortful, cognitively taxing search tasks" (p. 361). [1] discovered in another study that interruptions had an impact on frustration and annoyance. People who are distracted are likely to give more negative ratings.

Every-day tasks that participants experience as easy and straightforward result in a high difference in completion time between distracted and non-distracted groups. Tests which require a higher cognitive load do not show that high difference. This might be a reason why many information retrieval and usability studies came up with no obvious differences between laboratories and remote settings: the cognitive load in these studies – especially for complex retrieval test – was so high that the impact of distraction on performance was not perceptible.

3 Research Design

Given that context, what is new about this experiment is that it uses findings from studies in computer science and psychology as the basis for a new experiment that is specific to digital libraries. This experiment employs a remote setting that has not artificially eliminated external distractions. It measures the existence and impact of distraction in a remote setting that explicitly allows users to work in their own natural environment and compares it to a standard laboratory setting. It also seeks to validate findings about the influence of distraction on user ratings.

Based on the results of [4], the time to complete a task is used as an indicator for the existence and the amount of distraction. Completion time is therefore only a tool to demonstrate distraction. At the end of the experiment, participants were asked to answer questions about their distraction level (did they have applications open, had they looked at the applications, have they talked to others or did they have day-dreams). In order to avoid social-desirability-responses, participants were told that honest answers are important for this research and that their answers had absolutely no consequences for them. These questions were meant to confirm and explain the indications of distraction based on completion time.

The sample for this study consisted of Library and Information Science Masters students at the Berlin School of Library and Information Science. All participants already had a BA in Library and Information Science. They were expected to know how to search in digital libraries. 23 females and 8 males between 21 and 46 years old (average age: 26.6 years) participated in the test.

An email invitation was send to two working groups within a class using the e-learning platform Moodle. The email told the participants that the aim of the study was to test the usability of several digital libraries. They were explicitly told not to tell the others about the experiment, so that neither group would know that there was both a laboratory and a remote setting. The laboratory group got information about when to come to the laboratory where they would then get instructions; the natural environment group received a direct link to the test with instructions. These

instructions asked participants to do the test in their current environment: they were informed that they need not close any applications or to refrain from talking.

Both initial groups were similar in size, but not all students ultimately participated in the test. At the end, 13 participants completed the test in the laboratory (on a Monday afternoon at 2pm), 18 participants completed the test remotely in their natural environment at a time of their choosing. There was no significant relation between the hour when participants in the natural environment group completed the test – morning, afternoon or midnight – and the total time in seconds that they spent on the test. The test took place in November 2010.

The two groups underwent an asynchronous remote usability test using the software Loop11®. Each test was accessible by means of the Internet; the laboratory consisted of a computer pool. A moderator ensured that there was no external distraction and was sitting at the front desk. The situation resembled an exam.

The participants were asked to do similar tasks in five different digital libraries. In each digital library, users had to search for a specific document. The task description was shown on top of the screen; the system being tested was shown on the rest of the screen. Participants did not have to switch between windows to see the questions and the digital libraries. There was also no need to install any additional software. All digital libraries were fully functional within the test. The tasks included queries such as a search for the full text of the English version of Antigone by Sophocles or a talk that is entitled "Demokratie durch Krieg". Participants' confidence with the kind of digital libraries and the types of tasks was high, because the tasks resembled their preparations for essays or class papers. After each task, the users rated the degree of the task's difficulty.

The five digital libraries included three German ones – DigiZeitschriften (DL 1), Social Science Open Access Repository (DL 3) and Open Repository Kassel (DL 4) – as well as two English ones – Perseus Digital Library (DL 2) and Valley of the Shadow (DL 5). DL 4 and DL 5 were chosen because the tasks required a higher cognitive load compared to the other tasks. The aim of the latter was to test if the assumption is accurate that higher cognitively-loaded tasks result in no obvious difference between distracted and non-distracted participants.

All statements in the following result section about completion times and task ratings should only be seen in relation to the two settings: this research does not seek to demonstrate that people need a certain amount of time to complete a task in a specific digital library. The important information for this research is whether participants need more or less time in one of the two settings. The research's goal is not to analyze or to compare the usability of the five digital libraries – even if participants believe that it is.

4 Research Results

The three research hypotheses for the experiment are:

1. Participants in the laboratory should be faster, because their concentration is higher; participants in their natural environment should be relatively slower. This hypothesis is based on the results by [4].

2. Participants in the laboratory should rate the digital libraries more positively (that means they have a more positive cognitive reaction to the sites), compared to participants in their natural environment, who should give more negative ratings. This hypothesis is based on the results by [9].
3. External distractions should be a significant factor during test completion for the natural environment group. This will be validated as a new result not explicitly demonstrated in the scholarly literature.

In the following sections, the results based on the hypotheses are presented.

4.1 Hypothesis 1: Differences in Completion Times

The experiment affirms hypothesis 1: there are several differences between the two settings in the amount of time that participants needed. The mean time to complete the test in laboratory was 665.15 seconds; in the natural environment it was 1173.44 seconds, which is a clear indication of distraction in the latter.

An independent-samples t-test was conducted to compare the total completion time spent on the test by participants in the laboratory and in the natural environment. The data violates the assumption of equal variance between the two subject groups therefore equal variance is not assumed. There was a significant difference in scores for laboratory ($M = 665.15$, $SD = 153.73$) and natural environment participants ($M = 1173.44$, $SD = 845.01$; $t(18.54) = -2.50$, $p < .03$, two-tailed). Using the same test to compare the average time spent per task by participants in laboratory and in the natural environment also revealed a difference with $p < .03$.

There was no significant difference in scores for the number of clicks to complete the test within laboratory ($M = 28.85$, $SD = 8.30$) and natural environment ($M = 37.11$, $SD = 14.45$; $t(29) = 1.62$; $p > .10$, two-tailed). This means that although people in the natural environment needed statistically more time to complete the test than their comparable group in the laboratory (total completion time as well as average time per task), they did not have more trouble searching the documents. The number of clicks to complete the tasks is in statistical terms nearly the same. The reason for the longer test duration must lie elsewhere.

An additional factor between the two groups that has not come up in the literature yet is the high variance between participants in the two settings in the total completion time. This score measured in seconds how long participants needed to complete the whole test. In the laboratory, there is no significant difference in research terms between the five lowest scores (from 456 seconds for the whole test to 607 seconds) and the five highest scores (ranging from 682 seconds to 966 seconds). The natural environment group reveals a significant difference: the lowest score (473 seconds for the whole test) is as low as the one in the laboratory; but between the three highest scores in the natural environment are more than 1000 seconds (the three participants that needed the longest took 1593 seconds, 2731 seconds and 3693 seconds.) The laboratory shows an ideal user's behavior, whereas the behavior in the natural environment fluctuates heavily. Data with a high variance results in different interpretations than laboratory data that is normally distributed.

4.2 Hypothesis 2: Differences in Ratings

As assumed, the rating in the laboratory tends to be more positive with more participants rating tasks as easy (see table 1). This is a standard way of measuring positive or negative cognitive reactions among tasks as it is used in usability evaluation. In the natural environment group, more participants rated tasks negatively, that means they regarded more tasks as difficult or as neither difficult nor easy. This coincides with findings by [9].

Table 1. Average percentages of participants rating the 5 digital libraries in the laboratory (LAB) and in the natural environment (NE)

Options	Average rating of DL 1 to DL 5 in %	
	LAB	NE
easy	70.8	66.7
difficult	15.4	21.1
neither difficult nor easy	13.8	12.2

An interesting observation is that these differences are less obvious for the English language digital libraries DL 2 and DL 5 (a foreign language for the participants). Language may be another factor that influences user's behavior and will be examined in future studies. It is intriguing that the phenomenon described by [4] is mirrored in the results: for the most cognitively loaded tasks (here DL 4 and DL 5), the difference between the two groups becomes minor.

4.3 Hypothesis 3: Evidence of External Distraction in the Natural Environment

The experiment also affirms hypothesis 3: People in the natural environment group were distracted during the test. Only 22.2% said that they had no applications open during the test. 77.8% of the participants in the natural environment had applications open; of these, only 33.3% said that they never looked at the open applications during the test. Another 38.9% of participants said that someone had talked to them during the test (also via SMS or phone); 61.1% claimed that they talked to no one. One question for both setting groups was whether they had been distracted by day dreams such as thinking about their shopping lists or class preparations: 25.8% said no, while 74.2% admitted that they had been distracted by day dreams. About half of the whole sample said that they focused with 90% to a 100% of their attention on the test.

Is time really a good indicator for distraction and is distraction the reason for higher time rates in the natural environment group? A number of correlations with participants' answers to the additional questions about external distraction factors confirm this. The following numbers can only be seen as indications, because for most of the statistical tests only the small sample of the natural environment group (n=18) was considered. In the laboratory, nobody talked to participants and they were not allowed to have applications open during the test, so this group was excluded.

The relationship between having additional applications open during the experiment (as measured by the categorical variable [yes/no]) and the total time spent on the test (as measured by the continuous variable total time in seconds) was investigated using Pearsons coefficient. There was only a small, negative correlation between the two variables "open applications" and "time": $r = -.03$, $n = 18$, $p < .0005$. In other words: if applications are open, the total time tends very slightly to be higher. A second test shows that the more applications that are open, the higher the time rate.

The relationship between being talked to (as measured by the categorical variable [yes/no]) and the total time spent on the test was also investigated using Pearsons coefficient. There was a large, negative correlation between the two variables, $r = .56$, $n = 18$, $p < .0005$ with nobody talking to a participant, the smaller the time rates.

A frequency table showed that the percentage having day dreams was higher in the natural environment than in the laboratory: 89% of natural environment participants had day dreams and only 53.8% had them in the laboratory.

5 Conclusion

These results have several implications for user experience research, especially on the design of digital libraries. If researchers do tests in a laboratory, they should be aware that the natural environment is different. For example positive ratings may actually be less accurate than the numbers suggest.

This experiment showed that there are some statistically significant differences between participants in a laboratory and in the natural environment. These differences occur in the total time and the average time spent on the tasks and in the task's rating. There is no difference in the number of clicks per task. This means that participants in the natural environment executed the tasks similarly, even if they took longer and had different reactions about how easy they were. This phenomenon indicates that the difference in the completion time rates must have another source.

The indicator "completion time" is closely related to external distraction factors: participants who had admitted that they had looked at open applications talked to others or had day dreams also had a higher completion time score. The more that applications were open and the more that participants actually looked at these applications, the higher was the completion time. The interruptions with the highest impact were talks to others, either face to face or via phone.

If participants in a natural environment are distracted, designers should take it into account for their research design and analysis. For example, time-outs in the retrieval process can become a real barrier if a user is in a chat and wants to continue a search afterwards. Users are also less concentrated during the search process and will likely miss important information. Distracted people require repetitions of what they saw on an earlier page, because they will likely not remember it.

The actual use of digital libraries takes place in the natural environment. If researchers want to collect realistic data for user experience research, this experiment shows that the distraction-rich nature of the natural environment needs to be taken into account.

References

1. Adamczyk, P.D., Bailey, B.P.: If not now, when? The effects of interruption at different moments within task execution. In: CHI 2004, pp. 271–278. ACM Press, New York (2004)
2. Andrzejczak, C., Liu, D.: The effect of testing location on usability testing performance, participant stress levels, and subjective testing experience. Journal of Systems and Software 83(7), 1258–1266 (2010)
3. Bruun, A., Gull, P., Hofmeister, L., Stage, J.: Let your users do the testing: A comparison of three remote asynchronous usability testing methods. In: CHI 2009, pp. 1619–1628. ACM, New York (2009)
4. Czerwinski, M., Cutrell, E., Horvitz, E.: Instant Messaging and Interruption: Influence of Task Type on Performance. In: Australasian Computer-Human Interaction Conference, pp. 356–361. Southern Cross University, Harbour (2000)
5. Kelly, D., Gyllstrom, K.: An Examination of Two Delivery Modes for Interactive Search System Experiments: Remote and Laboratory. In: CHI 2011, pp. 1531–1540. ACM, New York (2011)
6. Huang, S.C., Bias, R.G., Payne, T.L., Rogers, J.B.: Remote usability testing: a practice. In: Proceedings of the 2009 ACM/IEEE Joint Conference on Digital Libraries, p. 397. Association for Computing Machinery, New York (2009)
7. Rieh, S.Y.: On the Web at home: Information seeking and Web searching in the home environment. Journal of the American Society for Information Science and Technology 55(8), 743–753 (2004)
8. Roethlisberger, F.J., Dickson, W.J., Wright, H.A.: Management and the worker: An account of a research program conducted by the Western electric Company, Hawthorne Works, Chicago. Harvard Univ. Press, Cambridge (1975)
9. Tullis, T., Fleischman, S., McNult, M., Cianchette, C., Bergel, M.: An Empirical Comparison of Lab and Remote Usability Testing of Web Sites. In: Usability Professionals Association Conference (2007)

Search Behavior-Driven Training for Result Re-Ranking

Giorgos Giannopoulos[1,2,⋆], Theodore Dalamagas[2], and Timos Sellis[1,2]

[1] School of ECE, NTU Athens
[2] IMIS Institute, "Athena" Research Center

Abstract. In this paper we present a framework for improving the ranking learning process, taking into account the implicit search behaviors of users. Our approach is query-centric. That is, it examines the search behaviors induced by queries and groups together queries with similar such behaviors, forming *search behavior clusters*. Then, it trains multiple ranking functions, each one corresponding to one of these clusters. The trained models are finally combined to re-rank the results of each new query, taking into account the similarity of the query with each cluster. The main idea is that similar search behaviors can be detected and exploited for result re-ranking by analysing results into feature vectors, and clustering them. The experimental evaluation shows that our method improves the ranking quality of a state of the art ranking model.

1 Introduction

In this paper, we focus on result re-ranking. To the best of our knowledge, the majority of previous works aim either at building a search model per user or at building common search models for users with similar search interests. However, these approaches usually consider each user's search history as a whole, without analysing it into its inherent search behaviors. Even when user's search history is further analysed, it is only performed in terms of content.

So, even though previous approaches have proven to be effective, they consider only one aspect of the problem: similarity of searches and users based on content. The other aspect, which is not handled, regards the latent search behaviors expressed by users. Two users may present similar search behavior only in some search topics, but completely different behavior in other ones. Also, it is often the case that a user searches in completely diverse topics that may lead to different search behaviors for the same user. Finally, it is possible that two searches on completely different topics (in terms of content) form, in fact, one single search behavior (in terms of the characteristics of the clicked results) for a user. We next give a motivation example to describe the above problems.

⋆ This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 316–328, 2011.

A PhD student A would like to search for papers related to IR. She clicks on results from ACM Digital Library. A proper model for this search behavior trains a ranking function that favors results whose url contains the word "ACM". User A also searches for information about a new cellphone. She clicks on video results that present cellphone's functionality. A proper model for this search behavior trains a ranking function that favors results with videos. Also, the user exhibits exactly the same search behavior when searching for sports cars. On the other hand, a researcher B wants to search, not for papers, but for research projects on IR. She clicks on results from sites with project descriptions and calls. She also searches for cellphones, looking for the exact information as user A.

It is clear that training a single ranking function or a ranking function per user (or per group of similar users), beside being infeasible in real world scenarios due to the user disambiguation problem, does not capture the diversity in search topics and search behaviors of users. Moreover, we cannot deal with the above issues considering only content similarity. We suggest training ranking models which are *search behavior specific* and *user independent*. For instance, in the previous example, a proper training involves: two distinct "IR related" ranking models M_1 and M_2, one for each user, respectivelly, and a third, both "cellphone related" (for both users) and "sport cars related" (for user A) model M_3.

Our approach. In this work, we present a framework that captures latent search behaviors and exploits them to train multiple *behavior-driven* ranking models. The main idea of our method is that similar search behaviors can be detected (and exploited for result re-ranking) by analysing search results into feature vectors, and clustering them. We next give a brief description of our approach:

1. We cluster all training queries to obtain groups of queries (*search behavior clusters*), whose click data are expected to train similar ranking functions.
2. For each search behavior cluster, we train a different ranking function, using only the clickthrough data of the queries belonging to the cluster.
3. For each cluster, we define its *textual representation* in order to be able to calculate the textual similarity of each new query with each cluster.
4. We then exploit the multiple rankings from step 2 and the query-cluster similarity scores from step 3 to produce a final ranking for each new query.

We built our method on top of Ranking SVM, a widely used machine learning technique. However, it is general enough to adapt to any other similar technique. Also, since our method works mainly on the training level, i.e., it improves the baseline training of a ranking model, it can be easily integrated with any of the already presented approaches for web search personalization. Finally, the proposed method fits every search scenario: from generalized web search to specialized search on Digital Libraries. We perform experiments showing that our method improves the ranking quality of a state of the art ranking model (RSVM).

The remaining paper is organized as follows. In Section 2, we discuss some background information on Ranking SVM that provides an intuition on our proposed solution. In Section 3, we present our method. In Section 4, we present the experimental results. Section 5 presents the related work and, finally, Section 6 conludes and discusses further work.

2 Background and Method Intuition

The example of Figure 1 gives a geometric interpretation of the above problem. For simplicity, consider that the feature vectors of the training data include only two features. Let feature x be the frequency of the word "acm" in the result text and feature y be the frequency of the word "video". The shaded shapes represent results related to paper searches, while the non-shaded shapes represent results related to cellphone searches. Squares correspond to strongly relevant results, triangles to partially relevant results and circles to irrelevant results.

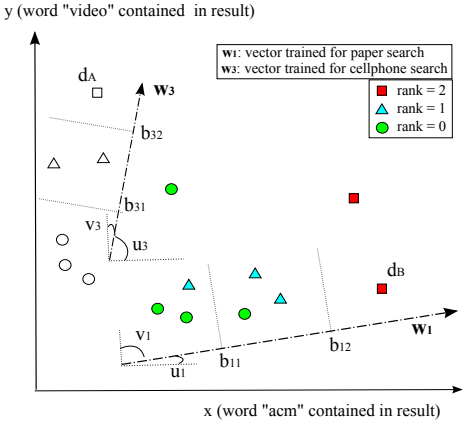


Fig. 1. Trained weight vectors and hyperplanes on feature space

In this feature space, we train two ranking models expressed by weight vectors $\mathbf{w}_1$ and $\mathbf{w}_3$. These vectors correspond to searches for papers and searches for cellphones, respectively. The direction (or *slope*) of each $\mathbf{w}_i$, i.e., the angle between the vector and one of the axes, indicates how important each training feature is for the process of result ranking. For example, the angle u_1 between vector $\mathbf{w}_1$ and x-axis is smaller than the angle v_1 between the vector and y-axis. This means that a change in the value of feature x is more probable to induce a change in the result's rank than a change in the value of feature y. So, for the particular training performed on clickthrough data from paper searches, feature x is more important than y when ranking query results. The opposite stands when training a vector $\mathbf{w}_3$ on cellphone searches: feature y is more important than x, as shown by the direction of $\mathbf{w}_3$.

The above example describes two different search behaviors, that is, specific search patterns followed by users for specific categories of searches. We can see that search behaviors are not expressed in terms of content, but through the feature space $\mathbf{X} \epsilon R^d$ selected to represent the clickthrough data (query results and their ranks/ relevance judgments). So, we can capture and exploit such search behaviors by utilizing the distribution of the training clickthrough data in the feature space. Next section describes our method.

3 Behavior-Driven Training

3.1 Query Clustering

The first step is to partition the initial training dataset into groups of queries whose clickthrough data are expected to train similar ranking functions, and thus, correspond to similar search behaviors, (*search behavior clusters*).

So, instead of running one Ranking SVM for each query, we approximate the ranking model to be trained on the query's clickthrough data. To this end, we define two categories of clustering dimensions related to the geometric characteristics of the model: *slope dimensions*, that correspond to the direction of the vector $\mathbf{w}$, and *margin dimensions*, that correspond to the points where $\mathbf{w}$ cuts through the rank-separating hyperplanes.

Dimension Selection for Clustering. As stated above, we define the following two categories of clustering dimensions: *slope* and *margin* dimensions.
Slope dimensions (f^g). With these dimensions we approximate the slope of the vector to be trained by the SVM model for each query. To extract them, we exploit the features from feature space $\mathbf{X}\epsilon\mathbb{R}^d$ that are used to represent query results. Specifically, we approximate the slope of vector $\mathbf{w}$ for each pair of features (x, y) where x, y features of $\mathbf{X}$. As shown in Figure 2, this is equivalent to approximating angle u.

The approximation is performed as follows. For each result rank r, we calculate the mean feature value of the results for each of the two features (x, y). So:

$$M^y_{r=2} = \frac{1}{m}\sum_{i=1}^{m} f^y_{i2} \tag{1}$$

for example, is the y coordinate of point M_{r2} (Figure 2), which is the centroid of all m results of rank 2 (square shapes) in the 2-dimensional feature space xy. f^d_{ir} denotes the value of feature d for the i^{th} result belonging to rank r.

Calculating the above expression for all ranks for both features x, y, we are able to approximate the tangent of angle u with the following formula:

$$f^g_{(x,y),(a,b)} = \frac{M^y_{r=b} - M^y_{r=a} + \varepsilon}{M^x_{r=b} - M^x_{r=a} + \varepsilon} \tag{2}$$

where (a, b) are rank pairs ($a, b \in \{0, 1, 2\}$ and $a \neq b$), (x, y) are feature pairs, and $\varepsilon = 10^{-9}$ is used to avoid zero-valued features and divisons with zero.

Each of the terms $f^g_{(x,y),(a,b)}$, calculated for all pairs of features (x, y), and for all pairs of ranks, is used as a clustering dimension, called *slope dimension*. So, for a feature space of size d and for ρ distinct ranks, we produce $\frac{d(d-1)}{2} \cdot \frac{\rho(\rho-1)}{2}$ slope clustering dimensions.

Note, however, that in most cases, the clickthrough data of each query includes much more results judged as irrelevant than partially or strongly relevant. Thus, we can extend Equation 2 by grouping all positive ranks into a single rank against to the rank of irrelevant results:

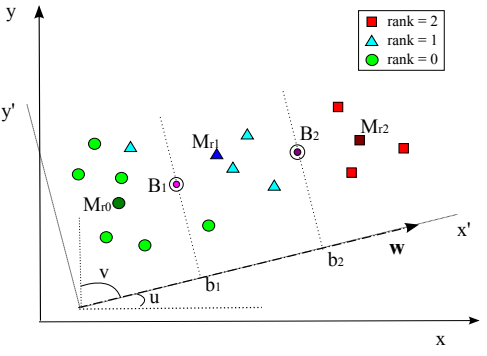


Fig. 2. Extracting clustering dimensions on feature space

$$f^g_{(x,y),(0,P)} = \frac{M^y_{r=P} - M^y_{r=0} + \varepsilon}{M^x_{r=P} - M^x_{r=0} + \varepsilon} \tag{3}$$

where $(0, P)$, $P \in \{1, 2\}$, is the rank pair of irrelevant and non-irrelevant results (partially or strongly relevant in our case). This produces a total of $\frac{d(d-1)}{2}$ slope clustering dimensions. The latter formula was finally prefered over Equation 2 since it gave slightly better results in the experimental evaluation.

Margin dimensions (f^m). With these dimensions we approximate the positions of points b_i on $\mathbf{w}$. These points define the normal hyperplanes which separate the results of different ranks. To approximate the positions of b_i, we use the projection of points B_i on $\mathbf{w}$, where B_i is the centroid of the result points of two neighbouring ranks. The coordinates of points B_i are calculated utilizing M (Equation 1) points, since B_i points are the centroids of neighboring M points. For example, in Figure 2,

$$B^y_2 = \frac{M^y_{r=2} + M^y_{r=1}}{2} \tag{4}$$

is the y coordinate of point B_2, which is the centroid of all results of rank 2 (squares) **and** rank 1 (triangles) in the 2-dimensional feature space xy.

In order to calculate the projection of B_i points on $\mathbf{w}$, we rotate the 2-dimensional space xy so that one of the two axes becomes parallel to $\mathbf{w}$. Without loss of generality, we rotate the space u degrees anticlockwise, where $u = \arctan f^g$ is approximated by the slope clustering dimensions. In the new space $x'y'$, $\mathbf{w}$ is parallel to x'-axis, so $b^{x'}_i = B^{x'}_i$, since segment b_iB_i is normal to $\mathbf{w}$.

So, finally, the margin clustering dimensions are calculated by the formula:

$$f^m_{(x,y),(a,b)} = \frac{\frac{1}{m}\sum_{i=1}^{m} f^{x'}_{ib} + \frac{1}{n}\sum_{i=1}^{n} f^{x'}_{ia}}{2} \tag{5}$$

where (a, b) and (x, y) represent the rank and feature pairs, respectively, $f^{x'}_{ir} = f^x_{ir}\cos u - f^y_{ir}\sin u$ are the feature values on the rotated axis x' and $u = \arctan f^g$ the rotation angle.

Similarly to the calculation of slope features, we can extend Equation 6 by grouping all positive ranks into a single rank:

$$f^m_{(x,y),(0,P)} = \frac{\frac{1}{m}\sum_{i=1}^{m} f^{x'}_{iP} + \frac{1}{n}\sum_{i=1}^{n} f^{x'}_{i0}}{2} \quad (6)$$

where $(0, P)$ is the rank pair of irrelevant and non-irrelevant results. This produces a total of $\frac{d(d-1)}{2}$ margin clustering dimensions.

The next step is to select a clustering methodology which exploits properly the extracted clustering dimensions to group search behaviors.

Clustering Methodology. We selected a partitional clustering method that utilizes repeated bisections [19], based on its reported effectiveness on datasets of similar form to ours. However, we note that there is probably room for improvement for this part of our method in future work. The method works as follows: All items (i.e., queries) are initially partitioned (bisected) into two clusters. Then, one of these clusters is selected and is further bisected. This process is repeated until we get the desired number of clusters. In each step, the selection of the cluster to be bisected and the bisection itself, is done in such a way that the bisection optimizes the value of a clustering criterion function.

The proper criterion function that leads to *center-based clusters*[1] aims at maximizing the following quantity:

$$\sum_{r=1}^{k} \sum_{d_i \epsilon S_r} \cos(d_i, C_r)$$

where k is the number of clusters, S_r is the set of items of cluster r, d_i the i^{th} item of cluster, C_r the centroid of the cluster and $\cos(d_i, C_r)$ the cosine similarity between each item of the cluster and the cluster centroid.

3.2 Ranking Function Training

For every extracted behavior cluster C_i, we train a different ranking function F_i using the Ranking SVM based on Joachims' model [1], [3]. As stated previously, each ranking function is trained using only clickthrough data from queries belonging to the corresponding cluster C_i.

The training features are the ones used in LETOR benchmark[2] and are described in [2]. Feature examples include TF, IDF, LMIR and BM25 considering, result title, abstract, body, url and pagerank values.

3.3 Matching Behavior Clusters with Queries

For each cluster, we extract the text of all its queries to define its textual representation. We consider each cluster's textual representation as a separate

[1] This cluster category fits best our case, where we practically cluster vector slopes and points on them, so items of the same cluster have to be closer to its centroid.

[2] http://research.microsoft.com/en-us/um/beijing/ projects/letor/LETOR4.0/Data/Features_in_LETOR4.pdf

document. We index these documents using the Lucene[3] IR engine. Given a query consisting of terms $t_1, t_2, \ldots, t_n$, we transform it into the following: (t_1 AND t_2 AND ... AND t_n) OR t_1 OR t_2 ... OR t_n. OR predicates are used to relax the initial query in order to find not only documents having all the terms but also documents having some terms of the query. This formulation also ensures that documents containing all query terms will get much higher score.

Then, we use Lucene scoring function to return a list of clusters, along with a score that indicates how similar to the query they are. We select the top-k most similar clusters that will be utilized in the next phases and discard the remaining. For the selected clusters, we normalize their scores so that the sum of those scores equals to 1. In the end, each new query q is assigned k weights w_i, $1 \leq i \leq k$, that represent its normalized matching scores (in terms of content similarity) with its k most similar search behavior clusters.

3.4 Re-Ranking

When a user poses a new query, its similarity score w_i with every behavior cluster C_i is calculated as described in Subsection 3.3. We consider the top-k most similar clusters for the remaining of the re-ranking process.

Then, using each ranking model F_i trained for each cluster C_i, we produce k different rankings R_i for the query, with rs_{ij} being the ranking score of result j according to model F_i. We should note here that Ranking SVM does not provide actual score values. Rather, it provides values that define the relative ranks of the results. In order to obtain a ranking score, we first define

$$rs_{ij} = M - r_{ij}$$

where M the number of returned results and r_{ij} the rank (position) of result j, according to model F_i, for the query. Then, the final score for each result j is:

$$score(j) = \sum_{i=1}^{k} w_i rs_{ij} \tag{7}$$

That is, we merge the different rankings produced by the ranking models that correspond to the most similar behavior clusters to the query.

However, our ranking function (Eq. 7) works better on high positions of ranking, while the baseline ranking (i.e. training a single ranking function on all clickthrough data), works better on lower rank results (see Section 4.2). Thus, for our final ranking function, we choose to favor our ranking score when scoring results at the top ranks, and the baseline score when scoring for lower ranks:

$$score(j) = \begin{cases} l \cdot \sum_{i=1}^{k} w_i rs_{ij} + (1-l) \cdot rs_j^b, \\ \text{for the top-a results according to the Eq. 7} \\ (1-l) \cdot \sum_{i=1}^{k} w_i rs_{ij} + l \cdot rs_j^b, \\ \text{for the rest of the results} \end{cases} \tag{8}$$

[3] http://lucene.apache.org/

where $l \epsilon [0.5, 1]$ weights the two rankings, α is the (ranking position) threshold at which the ranking weights are reversed and rs_j^b is the ranking score given by the baseline ranking function. The optimal values for α and l are computed after tuning on the training set (Section 4.2).

4 Experimental Evaluation

In this section, we present the experimental evaluation of our method and validate its effectiveness. Specifically, in Section 4.1 we present the experimental dataset and we describe the performed preprocessing. In Section 4.2, we compare our method with the baseline and in Section 4.3 we discuss the results.

4.1 Dataset Presentation and Pre-processing

For the experimental evaluation, we used LETOR [2], a benchmark dataset for research on learning methods for ranking search results. LETOR's latest version[4] uses the Gov2 web page collection and two query sets from Million Query track of TREC 2007 and TREC 2008. We focused on the most recent available dataset.

The initial benchmark dataset contains 784 queries with a total of 15211 judged results. Three labels (ranks) are used to judge the relevance of results: 0, 1, 2 (*irrelevant*, *partially relevant* and *strongly relevant* respectively). The query results are represented as feature vectors consisting of the training features mentioned in Section 3.2 and described in [2]. In LETOR, data is partitioned in five subsets. Combining each time different subsets to make the training, the validation and the test set, the LETOR authors create 5 different arrangements for five-fold cross validation. After removing queries with only irrelevant results (judgment 0), that are not expected to contribute to the training phase we resulted to five folds shown in Table 1.

Table 1. LETOR Dataset partition: number of queries contained in each set

Folds	Training set	Validation set	Test set
Fold1	339	119	105
Fold2	353	105	105
Fold3	347	105	111
Fold4	330	111	122
Fold5	322	122	119

4.2 Method Evaluation

In this section we compare our method, denoted M, to the baseline method of applying a single RSVM training for all queries, denoted B. The comparison is performed in terms of *Precision at position n (P@n)*, *Mean average precision (MAP)*, *Normalized discounted cumulative gain (NDCG)* and *Mean NDCG*.

[4] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4dataset.aspx

Comparison Process. The instance of Ranking SVM adopted by the LETOR benchmark[5] requires only one parameter c as input. LETOR sets c to the following values: {0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10}.

On the other hand, our method introduces four more parameters: N, the number of search behavior clusters to be created, k, the number of top clusters used for re-ranking, l, the weight of the ranking score of our method against the weight $(1 - l)$ of the score of the baseline method B (Eq. 8), and α, the offset at which the above weights are reverted.

We performed the evaluation as follows. We ran method B on each fold, for every value of c. For each fold, we tuned c to the value that results to the max MAP. We, also, ran our method M on each fold to tune c, N, k, l and α. First, we tuned c and N, varying the values of parameter N from 5 to 50 with step 5. Then, we tuned k, varying its value from 1 to 10, with step 1. Finally, we tuned l, varying its value from 0.5 to 1.0 with step 0.1, and a, varying its value from 1 to 10 with step 1. Note that we tuned the parameters to the values that result to the max MAP. To compare our method M with the baseline method B, we ran both methods using the tuned paramaters (and, thus, achieving the maximun MAP values).

The results presented next are obtained using the following parameter setting for our method: $k = 3$, $l = 0.8$ and $\alpha = 4$. The values of parameters N and c are presented for each Fold, separately, in Table 2.

Table 2. Optimal N and c parameters for each fold

	Fold1		Fold2		Fold3		Fold4		Fold5	
Method	**N**	**c**	**N**	**c**	**N**	**c**	**N**	**c**	**N**	**c**
B	-	2	-	0.2	-	0.2	-	1	-	5
M	10	2	25	0.1	5	0.5	30	1	20	10

Precision Results. Table 3 presents the *Mean Average Precision* and *Mean NDCG* values of our method M compared to the baseline B. As we can see, our method outperforms the baseline on both measures in all folds, except for *Mean NDCG* values for Folds 1, 3 and 5, where M and B have similar performance. The average MAP value on all folds is 2% higher than the baseline. This is an important improvement considering that we deal with the specific type of dataset. In such dataset types, users are asked to explicitly judge results for a given set of queries, and not to perform "real" search sessions that facilitate the extraction of search behaviors. However, as shown in the results, our method can exploit latent search behaviors to increase the precision of the training model.

Table 4 presents the $P@n$ values of our method compared to the baseline for each fold. As we can see, our method gives better precision for high rank results. Especially for $P@1$, our method achieves a max of 6% increase (4% average) on precision. This behavior can be explained as follows: for a large number of

[5] http://research.microsoft.com/en-us/um/beijing/projects/letor/letor4baseline.aspx

Table 3. Mean Average Precision and Mean NDCG

	MAP		Mean NDCG	
Folds	**B**	**M**	**B**	**M**
Fold1	0.68	0.70	0.70	0.70
Fold2	0.65	0.67	0.65	0.66
Fold3	0.64	0.65	0.66	0.66
Fold4	0.69	0.72	0.70	0.73
Fold5	0.66	0.67	0.68	0.68
Average	0.66	0.68	0.68	0.69

queries, each query is matched with high score with a certain behavior cluster (and lower scores with several others) which produces a model that favors the single most relevant query results.

In low rank results, our method has similar performance with the baseline. This is due to the low number of positive relevance judgments in the dataset, which restricts the precision scores on lower ranks.

Table 4. Precision Results

	Fold1		Fold2		Fold3		Fold4		Fold5		Av. Folds		Max Increase
Precision	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**M-B**
P@1	0.61	0.67	0.61	0.67	0.59	0.61	0.70	0.72	0.62	0.62	0.62	0.66	+0.06
P@2	0.60	0.60	0.52	0.54	0.53	0.54	0.67	0.67	0.58	0.58	0.58	0.58	+0.02
P@3	0.57	0.57	0.51	0.50	0.50	0.53	0.60	0.62	0.53	0.54	0.54	0.55	+0.03
P@4	0.54	0.53	0.48	0.48	0.48	0.49	0.56	0.57	0.49	0.50	0.51	0.51	+0.01
P@5	0.50	0.50	0.45	0.45	0.46	0.45	0.52	0.53	0.46	0.46	0.48	0.48	+0.01

Table 5 presents the $NDCG@n$ values for the two methods. Again our method outperforms the baseline on high ranks, while it has the same behavior with the latter on lower ranks. However, looking at the "Av. Folds" columns, where the average values on all folds are presented, we observe that our method improves $NDCG@n$ for most positions. Since higher $NDCG$ values are obtained when the most relevant results are ranked on the top (followed by the less relevant and the irrelevant ones), the above observation enforces our claim that our method favors the most relevant results of each query.

Table 5. NDCG Results

	Fold1		Fold2		Fold3		Fold4		Fold5		Av. Folds		Max Increase
NDCG	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**B**	**M**	**M-B**
NDCG@1	0.54	0.57	0.51	0.57	0.51	0.54	0.57	0.60	0.52	0.52	0.53	0.56	+0.06
NDCG@2	0.60	0.59	0.52	0.55	0.52	0.52	0.61	0.64	0.57	0.57	0.57	0.57	+0.03
NDCG@3	0.63	0.63	0.57	0.58	0.57	0.58	0.63	0.68	0.61	0.60	0.60	0.61	+0.05
NDCG@4	0.64	0.64	0.60	0.61	0.60	0.61	0.67	0.69	0.64	0.64	0.63	0.64	+0.02
NDCG@5	0.68	0.67	0.61	0.62	0.62	0.63	0.68	0.71	0.67	0.67	0.65	0.66	+0.03

4.3 Discussion

To sum up, our method increases the *MAP* and the *Mean NDCG* values in almost all cases, compared to the baseline method. It also increases significantly

the $P@n$ and $NDCG@n$ values in high ranks, while it maintains the baseline scores in lower ranks. However, there are some drawbacks related to the dataset used, which, to our opinion, restrict the potential of the method:

1. The size of the dataset. As shown in Table 1, the training set varies from between 322 and 353 queries, along with their clickthrough data. This is a factor that influences the quality of the ranking function training process, with negative effect on the performance of our method.
2. The characteristics of the dataset. The specific dataset does not constist of real user search history (queries and clickthrough data). It is composed from a set of queries and results, explicitly judged on their relevance. So, it is evident that explicit search patterns, emerging from similar searches, of users do not exist. However, some latent search patterns (behaviors) are yet captured, resulting to increased precision compared to the baseline. We believe that this process would be much more effective when applied on real user search clickthrough data.
3. The lack of distinct users. If the queries of the dataset where assigned to specific (distinct) users, we would be able to further personalize the re-ranking process, by introducing the user-cluster similarity, along with the query-cluster similarity presented in Section 3.4. Then, results would be re-ranked taking into account, both query-induced search behaviors, and user-specific search needs.

5 Related Work

In [4] the author proposes a topic-based refinement of the PageRank algorithm that allows the offline computation of a fixed number of PageRank vectors corresponding to specific topic categories. These vectors are then used to bias the computation of each query's results list, based on the similarity of the query to each topic category. In [5], the problem of PageRank personalization is also handled, with emphasis on scaling. In [6] the authors utilize concept hierarchies, like ODP[6], to categorize queries and build user profiles. Then, they use collaborative filtering techniques to re-rank query results based on those profiles. Compared to the above, our method differs in that we do not construct different profiles for each user, neither we consider predefined topic hierarchies. Instead, we utilize the information from several users to create search behavior clusters, in which users participate. Also, our method is based on search behavior similarity and not only on content similarity.

In [7], [8], [9], [10] the authors build their models utilizing users' short or long search history or context (e.g. desktop documents). These models are essentially user profiles that are exploited to expand future queries and/or refine their results. Compared to our method, those approaches consider only content similarity and, also, they do not exploit collaborative information from all users. In [18], we presented a preliminary work on training multiple ranking functions, based, however, on content and not on search behavior.

[6] http://www.dmoz.org/

There are also approaches that modify state-of-the-art machine learning techniques to achieve better ranking results. In [17], the classical Ranking SVM algorithm is modified so that (a) errors on top rankings are reduced and (b) the importance of queries with fewer relevant documents for the training phase is increased. In [11] the authors propose to train multiple ranking functions defined, however by different ranks and not different behaviors. Those ranking functions are then aggregated into a final ranker. Compared to our work, these approaches do not take into account the inherent relations between queries and their clickthrough data.

6 Conclusion

In this paper, we presented a methodology for improving the quality of ranking function training by capturing and exploiting latent search behaviors. The main idea of our method is that search behaviors are not necessarily content-dependent, and that they can be utilized to train more effective ranking models. The method is general enough to be combined with other personalization techniques, as well as to be applied in every search scenario. The experimental evaluation demonstrates the effectiveness of our method.

Future work involves experimenting on the clustering dimensions construction process, the clustering algorithm and metrics and on the query-cluster matching and weighting process. For example, finding a way to approximate the support vectors and use them instead of all results for the clustering dimensions construction process could, theoretically, produce better results. Also, we intend to test our method and verify the statistical significance of our results on much larger real user search datasets, where the notion of search behavior is more meaningful, and perform a more thorough cluster analysis. Finally, we want to study how individual user search behaviors can be exploited in our framework, thus comparing and integrating our query-centric method with user-centric and content-centric approaches.

References

1. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference, pp. 133–142 (2002)
2. Qin, T., Liu, T.-Y., Xu, J., Li, H.: LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. Information Retrieval Journal (2010)
3. Radlinski, F., Joachims, T.: Query chains: Learning to rank from implicit feedback. In: Proceedings of the Eleventh ACM SIGKDD International Conference, pp. 239–248 (2005)
4. Haveliwala, T.-H.: Topic-sensitive PageRank. In: Proceedings of the 11th International Conference on World Wide Web, pp. 517–526 (2002)
5. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the 12th International Conference on World Wide Web, pp. 271–279 (2003)

6. Rohini, U., Ambati, V.: Improving Re-ranking of Search Results Using Collaborative Filtering. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 205–216. Springer, Heidelberg (2006)
7. Chirita, P.-A., Firan, C.-S., Nejdl, W.: Summarizing local context to personalize global web search. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 287–296 (2006)
8. Tan, B., Shen, X., Zhai, C.: Mining long-term search history to improve search accuracy. In: Proceedings of the 12th ACM SIGKDD International Conference, pp. 718–723 (2006)
9. Sugiyama, K., Hatano, K., Yoshikawa, M.: Adaptive web search based on user profile constructed without any effort from users. In: Proceedings of the 13th International Conference on World Wide Web, pp. 675–684 (2004)
10. Shen, X., Tan, B., Zhai, C.: Context-sensitive information retrieval using implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference, pp. 43–50 (2005)
11. Qin, T., Zhang, X.-D., Wang, D.-S., Liu, T.-Y., Lai, W., Li, H.: Ranking with multiple hyperplanes. In: Proceedings of the 30th Annual International ACM SIGIR Conference, pp. 279–286 (2007)
12. Teevan, J., Dumais, S.-T., Liebling, D.-J.: To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. In: Proceedings of the 31st Annual International ACM SIGIR Conference, pp. 163–170 (2008)
13. Dou, Z., Song, R., Wen, J.-R., Yuan, X.: Evaluating the Effectiveness of Personalized Web Search. IEEE Transactions on Knowledge and Data Engineering 21, 1178–1190 (2008)
14. Zheng, Z., Chen, K., Sun, G., Zha, H.: A regression framework for learning ranking functions using relative relevance judgments. In: Proceedings of the 30th Annual International ACM SIGIR Conference, pp. 287–294 (2007)
15. Kim, J.-W., Candan, K.-S.: Skip-and-prune: cosine-based top-k query processing for efficient context-sensitive document retrieval. In: Proceedings of the 35th SIGMOD International Conference, pp. 115–126 (2009)
16. Chu, W., Keerthi, S.-S.: Support Vector Ordinal Regression. Neural Computation 19, 792–815 (2007)
17. Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y., Hon, H.-W.: Adapting ranking svm to document retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference, pp. 186–193 (2006)
18. Giannopoulos, G., Dalamagas, T., Sellis, T.: Collaborative Ranking Function Training for Web Search Personalization. In: Proceedings of the 3rd International Workshop PersDB 2009 (2009)
19. Zhao, Y., Karypis, G.: Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learning 55(3), 311–331 (2004)

An Organizational Model for Digital Library Evaluation

Michael Khoo and Craig MacDonald

The iSchool, Drexel University, 3141 Chestnut Street
Philadelphia PA 19104, USA
{Michael.Khoo,cmm366}@drexel.edu

Abstract. Evaluation is a central digital library practice. It provides important data for managing digital libraries and informing strategic decision-making. Digital library evaluation and management are organizational as well as technical practices. What evaluation models can account for these organizational factors, in practice as well as in theory? To address these questions, this paper integrates two models, one from the organizational literature (Porter's value chain), and one from the evaluation literature (evaluation logic models), into a generic, flexible and extensible evaluation model that supports the goal-oriented evaluation and management of digital libraries in specific sociotechnical contexts. A case study is provided.

Keywords: evaluation, lightweight, logic model, management, organization, planning, value chain.

1 Introduction

Evaluation is an important practice for digital libraries. It provides data for managers and other stakeholders and informs long-term strategic decision-making [1]. This paper proposes an organizational model for evaluating and managing digital libraries. The approach is based on Porter's value chain model of organizations, and evaluation logic models, and focuses on evaluating digital libraries as sociotechnical phenomena that serve diverse communities, domains, and audiences. It addresses the needs of a range of stakeholders (managers, developers, users, funders, etc.) who may have limited resources (time, expertise, etc.) to carry out an evaluation [2].

2 Scope of the Research

A digital library can be defined in many different ways. Different definitions will shape evaluation approaches in different ways. The definition of a digital library that will be used in this paper is that of an operational organization that manages collections and services in multiple contexts of use. The term 'operational' is used here to indicate that a digital library has been released to users, is accessible on a regular basis, and is responsible to a range of stakeholders. Note that an operational digital library's staff might not necessarily have a background in the digital library research community, a situation that is increasingly the case as the technologies for

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 329–340, 2011.

building digital libraries become more widely available, and more organizations seek to provide searchable access to online collections. Operational digital libraries are sociotechnical systems, networks of technology, information, documents, people, and practices that interact in complex ways with wider worlds of work, institutions, knowledge, society, and knowledge production [3]. An evaluation model for a digital library should account for this complexity [4].

2.1 Existing Research

To situate the proposed organizational model of a digital library, this section reviews some existing digital library evaluation approaches (see also [5, 6]). Note that approaches that focus predominantly on usability (e.g. [7, 8]) are not covered.

Digital library evaluation approaches typically propose a holistic model of a digital library. Borgman et al. [9], addressing the social aspects of digital libraries, introduce the 'information lifecycle,' a holistic cycle of information use by individuals, organizations, and communities. Digital libraries are defined here as "computer-based systems constructed ... in a way that accommodates the actual tasks and activities that people engage in when they create, seek, and use information resources." This model it integrates a diverse range of digital library activities and can be used to generate use cases for different communities of users. The authors note the following evaluation research questions: "What kind of comprehensive measures do we need to design that evaluate the whole information and learning experience? What kind of evaluation processes (and supporting tools) will provide timely and valid predictions about individual steps, features, and capabilities?"

Saracevic [10] defines a digital library as "a set of elements in interaction" with one or more objectives, operating in a series of environments. Evaluation considers the effectiveness of system performance in relation to stated goals. It must address five criteria, the first of which is deciding what to evaluate - "what is meant by a digital library?" The rest of the evaluation criteria – the context of the evaluation (goals, levels of analysis, etc.), the performance criteria for selected objectives, specific measures for the criteria, and the methodology for carrying out the evaluation (instruments, procedures, etc.) – all flow from this definition.

Marchionini et al. [11] describe digital libraries as mixtures of complex human information seeking behaviors and rapidly evolving technologies, and advocate a user-centered approach to evaluation and design focused on "the information needs, characteristics, and contexts of the [users]." They outline an iterative human-centered evaluation process, in which needs are expressed in terms of a variety of goals, and may be evaluated in different ways, using different measures and methods, over different periods of time. This 'process-oriented' approach contains three principles: (1) know the users, (2) embed design and evaluation efforts into the organizational structure, and (3) create flexible systems and tools that can meet the requirements of people with diverse characteristics and needs. The approach is useful in early design stages, when evaluation targets can be aligned with operational goals.

Gonçalves et al. [12, 13] state that digital libraries are "extremely complex information systems … [and] the proper concept of a digital library … evades definitional consensus." Their '5S' model works towards establishing a consensus by describing five digital library components: Streams, Structures, Spaces, Scenarios,

and Societies. They specify a "minimal digital library" model consisting of digital objects, a metadata specification, a collection, a repository, and services. Evaluation with the '5S' model addresses quality dimensions for these components using established performance indicators (e.g., the metadata specification can be assessed for accuracy, completeness, and conformance).

Fuhr et al. [14] model digital libraries as complex systems that can be viewed from many perspectives. Their 'interaction triptych' model of digital libraries (c.f. [15]) identifies three major evaluation loci: usability, the quality of interaction between the user and the system (effectiveness, satisfaction, etc.); usefulness, the relevance of the information for the user; and performance, related to system performance attributes (precision, recall, response time, etc.). They recommend that evaluation should be flexible and extensible to cope with rapidly changing digital libraries; involve practitioners and real users; use standardized platforms for gathering, storing and disseminating evaluation data; evaluate user behavior in-the-large, along with sociological, business, institutional and other factors; and relate digital library models to models in other areas (archives, portals, knowledge bases, etc.).

The DELOS Digital Library Reference Model [16] defines digital libraries in terms three distinct 'layers': *organizations* that manage collections of digital content; *systems* that connect users to the library's collections; and *management systems* that manage user interactions with the library and collections. These layers support six core concepts: content (data and information made available to users): users (the various actors entitled to interact with a digital library); functionality (the services offered to different users); quality (the parameters for evaluating the content and behavior of a digital library); policy (the sets of conditions, rules, terms and regulations governing interaction between the digital library and users); and architecture (a mapping of the functionality and content offered by a digital library on to hardware and software components). The model can be expanded in a number of directions; for instance, users are defined in terms of end users, designers, system administrators and application developers.

The evaluation approaches discussed so far have mainly introduced *theoretical* models for digital library evaluation. The *practice* of digital library evaluation can be very complex, particularly for large-scale projects. Chowdhury et al. [6] report the evaluation of the Scottish Cultural Resources Access Network (SCRAN) to assess its value to the public libraries licensing it. The evaluation took into account a number of complex social, economic and political factors. A multi-method approach was deployed, including usability, interviews, web log analysis, surveys of library staff and users, and analysis of project minutes and documents. The analysis triangulated the factors affecting usage, and users' and library staff perceptions, with the aim of determining the overall value of SCRAN. The evaluation outcomes included recommendations for the management of similar projects such as the need for effective marketing, outreach, and rights management, and the question of whether a national cultural heritage information architecture should be centralized or distributed. The evaluation demonstrated how digital library operation and evaluation in real world settings can include complex management decisions.

2.2 The Organizational Value Chain

As the SCRAN example shows, there is a need for digital library evaluation models to address organizational and management factors. One influential model of a firm is Porter's 'value chain' (Figure 1). A value chain is "a collection of activities that are performed to design, produce, market, deliver and support its product" [17]. It is closely related to organizational structure, and organizations are often designed around these activities. The five primary activities outlined by Porter are: inbound logistics; operations; outbound logistics; marketing and sales; and service and support. These activities are represented in the figure as vertical columns from left to right. Note that there are downstream dependencies from left to right: for instance, the quality of inbound logistics will affect the quality of outputs, even if the two are not directly linked in the figure. There are also four support activities, represented as horizontal layers on top of the primary activities: infrastructure (general management, etc.); human resources; technology (R & D, new techniques, etc.); and raw material procurement. These activities contribute indirectly to production, but they are important for sustaining overall organizational processes; for instance, management information systems can play a crucial role in integrating the activities of a firm.

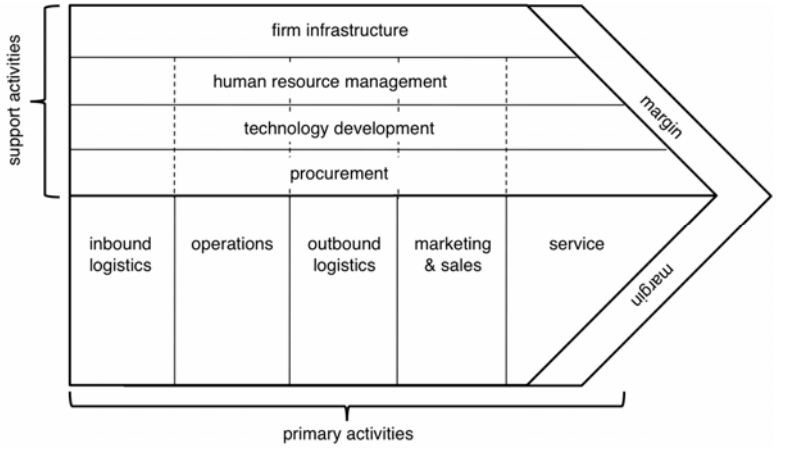


Fig. 1. Porter's Value Chain ([17] fig. 2-2)

2.3 Logic Models

Having established a value chain model of an organization, the next step is to think about how to evaluate such an organization. Logic models are a useful way to do this. Logic models are generalized, graphical models of an organization [18, 19, 20]. A simple logic model consists of four stages: (1) inputs, such as staff and funding, that support organizational activities; (2) the activities an organization engages in on a day-to-day basis that transform inputs into outputs; (3) outputs, that is, immediate products and benefits of organizational activities; and (4) long-term outcomes, such as changes in wider social contexts [21] (Figure 2).

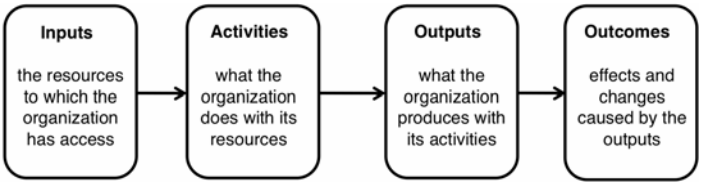


Fig. 2. A generic logic model

Given a set of inputs, organizational activities will produce outputs and outcomes. Logic models can identify evaluation areas in an organization by assigning the goals of an organization to the outcomes stage, and then working backwards through the model to identify the outputs, activities and inputs that support these goals and outcomes [22]. This method ensures that once activities are identified, they do in fact contribute to project goals and outcomes. Once a relevant set of organizational activities is identified, appropriate evaluation work for each activity can begin. This evaluation work should consider the nature of each activity and the appropriate benchmarks and techniques for evaluating it, as well as links to other activities, and the organizational resources (time, funds, expertise, etc.) available for the evaluation. A number of examples of logic model have been published [e.g. 23, 24]. There are a number of similarities between the value chain and the logic model. Both model organizations, and the inbound, operations and outbound stages of the value chain correlate very well with the inputs, activities, and outputs/outcomes stages of the logic model. What is missing from the logic model however is the infrastructural dimension of the value chain (the support activities).

2.4 An Organizational Model for Evaluating Digital Libraries

The two approaches just introduced are combined in this section to produce a generic organizational model, which can be used for evaluating digital libraries. This model includes the four stages of the evaluation logic model, and adds an additional organizational 'layer' from the value chain model. The example provided in Figure 3 models a digital library as a goal-oriented organization that receives inputs from the external environment, carries out transformational activities on those inputs, generates a series of outputs and outcomes, with these processes being coordinated through organizational knowledge and communication (documents, meetings, e-mail, etc.). In this case, the model is populated with goals and activities that are typical of some digital libraries today, but it could be repopulated with different goals and activities that reflect a different form of digital library. Working backwards, the *outcomes* stage focuses on the broader impacts of digital library use (such as better learning) and is defined first (for instance from mission statements). The *outputs* stage is designed to capture the major outputs required to support the wider outcomes (e.g. resources, collections, metadata, a web site, and a user community). The *activities* stage of the model is designed to list the activities that will be used to produce the outputs; note that the example activities listed in this stage – resource creation, metadata creation, web site, user services, and community building – themselves have internal linkages. (For the sake of this model, the linkages are as follows: resources are created, then

metadata is applied to resources, then resources are made available through a web site, then user services can be developed to support library use, then a user community can be developed to support users.) This implies further downstream linkages within the activities stage: for instance, poor resource and metadata quality can affect web site usability and community building. The *inputs* stage describes the general inputs such as an organizational mission, staff, equipment, and funding, that support overall work in the library. Finally, the *organizational* stage describes the organizational infrastructure that will support the first four activities.

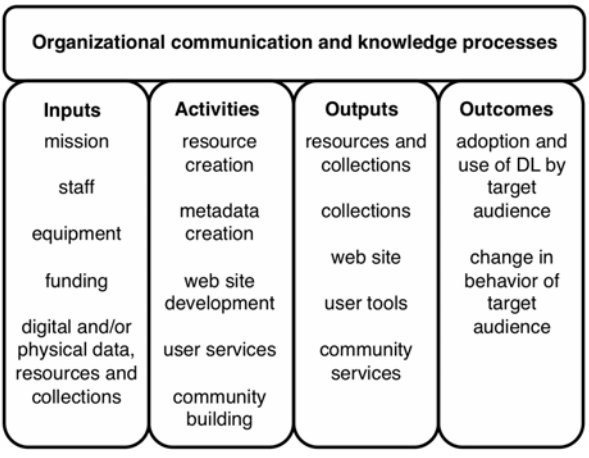


Fig. 3. An example organizational evaluation model for digital libraries

It is important to note that this particular description is just one example that can be generated by the model. As emphasized above, the model is generic, flexible extensible, and is capable of generating evaluation models for any type of current or potential digital library. While it happens to conform to some common current models of a digital library, a change in the outcomes stage could generate a completely different model. For instance, a library emphasizing the community design and sharing of interactive educational resources, supported by interactive repositories with tagging, folksonomies, etc., would lead to a different set of activities.

3 Evaluation Case Study

This case study describes the application of this evaluation model with the Internet Public Library ('ipl2': http://ipl.org). The main purpose is to illustrate the use of the evaluation model in the design and implementation of an evaluation plan, rather than to provide a detailed technical evaluation of ipl2. ipl2 was founded in 1995 and now has 40,000 online resources, 12,500 web pages, and a virtual reference service that has answered over 95,000 questions. In 2008, work began on a merger with another educational digital library, the Librarian's Internet Index (LII); this included crosswalking both ipl2 and LII metadata to Dublin Core, and merging their subject

browsing hierarchies. To carry out this work, ipl2 has two staff members and a voluntary team of about ten faculty, graduate students, IT staff, and a director. Developing an evaluation plan for ipl2 involved balancing evaluation goals against these limited resources and organizational constraints.

The process began by identifying ipl2 end users' needs. As with many digital libraries, this initial step was problematic, as it was not known who the ipl2 users were. Work therefore began with placing the ipl2 vision statement (Figure 4) in the outcomes stage of the model outlined above (Figure 3). Creative solutions to evaluating to what extent these goals were being met were then sought. Three initial evaluation areas were identified: community, usability and metadata. Examples of the evaluation work are given in the following paragraphs.

Vision Statement of ipl2:

ipl2 will shape and direct the evolving role of libraries in an increasingly digital world while working to become a virtual learning laboratory for the study of information services and technology.

The ipl2 staff, faculty, and volunteers will strive to meet these goals by:

1. creating and supporting new learning experiences for students and volunteers,
2. serving as a research forum and as a technological test-bed,
3. serving the public by finding, evaluating, selecting, organizing, describing, and creating high quality information resources, and
4. connecting members of the public to high quality information resources through the IPL question answering service.

Fig. 4. ipl2 mission and vision (http://www.ipl.org/div/about/mission_and_vision.html)

The ipl2 community. ipl2 relies on a large community of students and volunteers to maintain its digital collections and virtual reference service. The library serves as a teaching and learning environment for library and information science faculty and students at universities in the United States and Canada, and for professional librarians who volunteer their time and expertise. Feedback surveys from instructors and students using ipl2 were identified as a useful way to evaluate the community members who help to run ipl2. The results revealed that instructors and students see ipl2 as a useful learning site but that there are concerns with the quality of the tools, and the ipl2 collection and search results. In other work, a small sample (n = 528) of online reference answers was found to conform to ipl2 guidelines, suggesting that the information classes were themselves functional [25]. Future evaluations of community tools will include usability testing of a new metadata tool for volunteers, which will be deployed in various cataloging and metadata classes.

Web site usability. Usability was chosen as an evaluation activity because several faculty and students in the team had usability as a research and/or teaching interest. Beginning in 2008, the web site was completely redesigned, and testing of the new web site was carried out in several ways that drew on available resources. Student assistants on the project carried out paper prototyping, heuristic evaluation, and user testing of the new site. Approximately 100 students in graduate student 'Introduction to HCI' classes carried out heuristic evaluations of the ipl2 site as a class assignment. Positive findings included satisfaction with the site's 'look and feel,' layout, etc.

Negative findings focused on poor search experiences (e.g. lack of advanced search, lack of ranked search results), navigation, confusing browsing and subject headings, lack of help tools, etc. This usability testing provided data that triangulated with the community evaluations, particularly regarding efficacy of the search tools.

Metadata quality. ipl2 metadata has been created at different times and in different formats. Metadata evaluation work was constrained by these legacy issues, and by the complexities of the existing metadata. An audit of 15,000 MySQL records revealed approximately 53 different fields (including administrative fields) only a few of which were filled at or near 100% (e.g. *Title*, *Abstract*, and *URL*), while many fields (e.g. the former URL of a site) and were rarely used (c.f. [26]) (this evaluation supported the design of the new application profile and crosswalk to Dublin Core). An emergent issue concerned the browsing metadata. The project team wanted to generate a common browsing structure that described both ipl2 and LII resources in a common structure. Analysis of the existing browsing metadata showed however that the relationships between topics and subtopics in each library were complex and web-like in nature, making it impossible to generate a definitive hierarchical list. A pilot analysis of the LII for example identified over 42,000 separate parent-child relationships, and this issue has significantly slowed down metadata work in ipl2.

Overall, the metadata evaluation was lightweight and strategic rather than systematic and comprehensive, and driven by the fit with team members' research interests. Even these limited evaluations found significant issues that needed addressing [27], and which could be triangulated to other findings, such as the usability testing which suggested that it was hard to search and browse ipl2.

Web Metrics. Web metrics analyze visitor traffic to and through a web site; tools differ widely in functionality, complexity, and cost [28]. ipl2 uses Google Analytics (GA), a free javascript-based tool that reports data to external servers, from where it can be accessed through a browser via a user account [29]. GA required significant overhead to install but relatively little maintenance. Data showed that users spend more time viewing ipl2-created content than searching or browsing the collections, and that some of the popular collections were accessed largely from external search engines rather than the site's own search tool. This suggests that visitors were having trouble accessing these collections through the internal search, again triangulating with findings from other evaluation activities regarding search functionality.

Organizational processes. The final evaluation area covered ipl2's organizational knowledge and communication processes (a support activity rather than a primary activity - Figure 2). Interviews with project members identified several knowledge and communication issues in ipl2 arising from factors such as the long history of the IPL, the existence of a number of poorly documented legacy systems, an organizational history of voluntary and minimal staffing, and a voluntary group of project members with disparate backgrounds (IT, IS, libraries, digital libraries, HCI, etc.). It was difficult to coordinate ipl2 work across various technical and organizational issues, and project communication could be fragmented. This evaluation of organizational processes again pointed out the some of the underlying issues with ipl2 metadata, as project members commented on the difficulties involved in working with legacy metadata and understanding how earlier metadata had been

formatted, which in turn led to issues with designing new tools based on the metadata, such as the browsing tool.

Evaluation outcomes. As was emphasized in the introduction to this section, the purpose of the case study is not to provide a thorough technical evaluation of ipl2, *but to show how such an evaluation could be guided by the model introduced above*. At the beginning of the evaluation, the first and last stages of the model were populated by the existing (and very limited) organizational resources available to ipl2, and the overall goals of ipl2. The evaluation then became a question of how to bridge these stages with available organizational resources. It was decided to follow a 'good enough' strategy with selected activities, in the hope of generating enough data for useful triangulation. This strategy proved productive. Multiple evaluation activities produced data that indicated that search and browse were problematic, and there was evidence that organizational processes were linked to these problems. Work has since been initiated to improve the underlying catalog and database, as well as the associated tools and organizational processes.

While organizational factors were not directly identified in the ipl2 mission and vision statements, they did contribute to the overall goals; and deficiencies in these factors could therefore contribute to pl2 goals not being met. The evaluation results may therefore be framed not in the form of "ipl2 performed at X% on this measure," but rather in terms of "it appears (from a number of ongoing evaluation activities) that there is an issue with metadata and search, and organizational processes – what further work needs to be done?"

4 Discussion, Limitations, and Future Work

Evaluating ipl2 is a complex activity, especially for a largely voluntary and part-time organization with limited resources. The organizational evaluation model outlined above provided a useful way to describe ipl2's many components in terms of a small number of priority areas, which could be matched with available organizational resources. Evaluation efforts were targeted on feasible and practical evaluation tasks within each of these areas. A large amount of pragmatism was involved, with evaluation activities being carried out by project members with different skills and backgrounds, following different evaluation and research questions. A key outcome was that the model provided a useful way to understand the functional linkages between ipl2 activities, and to integrate and triangulate the results of the different evaluation activities. The evaluation results pinpointed areas where improvements should be made, especially with regard to metadata, search and site navigation, and also organizational communication and knowledge processes. The findings are being used to inform ongoing library development.

How do these findings compare with other evaluation models? Compared with the approaches summarized above, a number of components of the current model are familiar, for instance in Borgman et al.'s [7] 'lifecycle' model, and Marchionini et al.'s [25] multifaceted approach to user-centered design. An important contribution of the current model is that it adds a specific organizational dimension to evaluation; shows how this is related to technical dimensions; and shows how the organizational dimension can be implemented and assessed in an evaluation study.

An advantage of the model is that, being derived from business and organizational theory, and evaluation theory, it is theoretically independent of digital libraries. The approach is therefore flexible, extensible and scalable, and it can be adapted and applied to any type of digital library (current or yet to be built), at any scale, in order to generate a description of that digital library that can then be used as the basis for evaluation and management. The addition of the organizational dimension strengthens the ability of the model to link various activities together in a way that provides a bridge between the evaluation of small-scale and day-to-day aspects of library operations and the longer term and more open-ended organizational decisions and goals. The model provides a way to make those linkages – which are different in each digital library – explicit and observable in practical contexts. As Fuhr et al. [12] note, "The quality of a complex system is never better than the quality of its 'weakest' component," and evaluation must therefore take into account the nature of digital library components and the linkages between them.

The case study identified several limitations and opportunities for future research. First, the model pays attention to the organizational and managerial dimensions of operating and evaluating digital libraries, but, for reasons of space, it does not explore these dimensions in depth. There is an opportunity to develop the model to include other organizational and management theories, with the hope that these can provide further insight into the organizational nature of digital libraries. How, for instance, might models of organizational communication and knowledge be useful for digital libraries? Second, the model suggests that resource and metadata quality are key digital library activities that have a number of downstream linkages; what therefore are the organizational dimensions of digital libraries that can affect resource and metadata quality? Third, a disadvantage (shared by other models) is that evaluation work is often resource-intensive. There is therefore a need to develop lightweight and easy-to-implement evaluation techniques that can be used with the model, and for comparisons of these techniques that can guide a digital library *as an organization* to optimally allocate organizational resources. Finally, the framework recognizes the importance of, but does not substantively address, wider questions related to the long-term sociotechnical impact and outcomes of digital libraries (for example in terms of educational impact). Considerable work still remains to be done to develop evaluation models to investigate and address these impacts.

5 Conclusion

The digital library evaluation model introduced in this paper integrates concepts from the business and evaluation literatures to build a generic, flexible and extensible model for the goal-oriented evaluation and management of digital libraries. The framework is designed to support the optimal planning and allocation of evaluation resources within an organization. A case study of the use of the model showed that it provided coherence and focus for a range of evaluation practices in the context of a number of constraints, including limited organizational resources and complex legacy architectures. The model balanced evaluation requirements, constraints, and resources, and generated a shortlist of evaluation activities based on available resources, including the expertise of the project team. It supported the team to look for creative ways in which evaluation activities could

be folded into existing research and teaching activities. It provided a scaffold for triangulating disparate evaluation findings. The case study suggests that the model is useful and provides a way to gain traction with complex real world digital library evaluation issues. It will be of use to digital library managers, developers, and also external stakeholders such as higher-level administrators and funders.

References

[1] Reeves, T.C., Apedoe, X., Woo, Y.H.: Evaluating Digital Libraries: A User- Friendly Guide. National Science Digital Library, Boulder (2005)
[2] Khoo, M., MacArthur, D., Zia, L.: An agency perspective on digital library evaluation. In: Papatheodorou, C., Tsakonas, G. (eds.) Evaluation of Digital Libraries, pp. 41–59. Chandos Publishing, Oxford (2009)
[3] Bishop, A.P., Van House, N.A., Buttenfield, B.P. (eds.): Digital Library Use. Social Practice in Design and Evaluation. The MIT Press, Cambridge (2003)
[4] Khoo, M.: A Sociotechnical Framework for Evaluating a Large-Scale Distributed Educational Digital Library. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 449–452. Springer, Heidelberg (2006)
[5] Borgman, C.: Designing digital libraries for usability. In: Bishop, A.P., Van House, N.A., Buttenfield, B.P. (eds.) Digital Library Use, pp. 85–118. The MIT Press, Cambridge (2003)
[6] Chowdhury, G., McMenemy, D., Poulter, A.: Large-Scale Impact of Digital Library Services: Findings from a Major Evaluation of SCRAN. In: Gonzalo, J., Thanos, C., Verdejo, M.F., Carrasco, R.C. (eds.) ECDL 2006. LNCS, vol. 4172, pp. 256–266. Springer, Heidelberg (2006)
[7] Blandford, A., Adams, A., Attfield, S., Buchanan, G., Gow, J., Makri, S., et al.: The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work. Information Processing and Management 44(1), 4–21 (2008)
[8] Jeng, J.: Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. In: LIBRI, vol. 52(2-3), pp. 96–121 (2005a)
[9] Borgman, C.L., Bates, M.J., Bates, M.V., Efthimiadis, E.N., Gilliland-Swetland, A.J., Kafai, Y.B., et al.: Social aspects of digital libraries. Graduate School of Education & Information Studies, University of California, Los Angeles, CA (1996)
[10] Saracevic, T.: Digital library evaluation: Toward an evolution of concepts. Library Trends 49(2), 350–369 (2000)
[11] Marchionini, G., Plaisant, C., Komlodi, A.: The people in digital libraries: Multifaceted approaches to assessing needs and impact. In: Bishop, A., Buttenfield, B., VanHouse, N. (eds.) Digital Library Use, pp. 119–160. MIT Press, Cambridge (2003)
[12] Gonçalves, 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)
[13] Gonçalves, M.A., Moreira, B.L., Fox, E.A., Watson, L.T.: "What is a good digital library?" – A quality model for digital libraries. Information Processing and Management 43(5), 1416–1437 (2007)
[14] Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., et al.: Evaluation of digital libraries. International Journal on Digital Libraries 8(1), 21–38 (2007)

[15] Tsakonas, G., Kapidakis, S., Papatheodorou, C.: Evaluation of user interaction in digital libraries. In: The DELOS Workshop on the Evaluation of Digital Libraries, Padua, Italy (2004), http://dlib.ionio.gr/wp7/WS2004_Kapidakis.pdf (retrieved)
[16] Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., et al.: The Digital Library Reference Model. Foundations for Digital Libraries: DELOS Network of Excellence on Digital Libraries Project (2010)
[17] Porter, M.: Competitive advantage. Creating and sustaining superior performance. The Free Press, New York (1998)
[18] Khoo, M., Giersch, S.: Planning digital library evaluation with logic models. In: Papatheodorou, C., Tsakonas, G. (eds.) Evaluation of Digital Libraries, pp. 217–234. Chandos Publishing, Oxford (2009)
[19] Cooksy, L.J., Gill, P., Kelly, P.A.: The program logic model as an integrative framework for a multimethod evaluation. Evaluation and Program Planning 24, 119–128 (2001)
[20] Millar, A., Simeone, R.S., Carnevale, J.T.: Logic models: a systems tool for performance management. Evaluation and Program Planning 24, 73–81 (2001)
[21] McLaughlin, J.A., Jordan, G.B.: Logic models: a tool for telling your program's performance story. Evaluation and Program Planning 22, 65–72 (1999)
[22] Frechtling, J.: The 2002 User Friendly Handbook for Project Evaluation. The National Science Foundation, Arlington (2002)
[23] Centers for Disease Control and Prevention, Introduction to process evaluation in tobacco use prevention and control. U.S. Department of Health and Human Services, Atlanta(2008)
[24] Kellogg Foundation, Logic model development guide. W. K. Kellogg Foundationm, Battle Creek (2004)
[25] Burger, A., Park, J.-R., Lu, G.: Application of Reference Guidelines for Assessing the Quality of the Internet Public Library's Virtual Reference Services. Internet Reference Services Quarterly 15(4), 209–226 (2010)
[26] Mayernik, M.: The distributions of fields in bibliographic records. A Power Law Analysis. Library Resources and Technical Services 54(1), 40–54 (2009)
[27] Khoo, M., Hall, C.: Merging Metadata: A Sociotechnical Study of Crosswalking and Interoperability. In: JCDL 2010, Brisbane, Australia, pp. 361–364 (2010)
[28] Khoo, M., Pagano, J., Washington, A., Recker, M., Palmer, B., Donahue, R.: Using Web metrics to analyze digital libraries. In: JCDL 2008, Pittsburgh, PA, pp. 375–384 (2008)
[29] Google (n.d.). Google Analytics, http://code.google.com/apis/analytics/docs/

Developing National Digital Library of Albania for Pre-university Schools: A Case Study

Xiaohua Li[1] and Ardiana Sula[2]

[1] Sacred Heart University Library,
5151 Park Avenue, Fairfield CT, U.S.A
[2] Ministry of Science and Education
Rr. Durresit, Nr 23,
Tirana, Albania. AL -1001

Abstract. While the concept of digital library (DL) is well perceived and applied in developed countries, it is still a big challenge to the developing nations. There are great disparities, known as digital divide between developed countries and developing countries in terms of electronic resource funding, availability, and accessibility. DL, together with the information retrieval (IR) system, is believed to be an effective way to mend the gap of digital divide. This paper will employ a real case to discuss the significance of developing a national level of digital library for pre-university schools of Albania, the challenges of designing such information system both economically and technologically, and considerations of designing the digital library.

Keywords: DL, digital library development, developing countries, ICT infrastructure, IR, information retrieval, IS, information system.

1 Introduction

Albania is a developing country in Southeastern Europe with population of approximate 3.2 million. Large number of population (600,000) resides in the capital city, Tirana. ICT (Information and Communications Technology) education is not new to the nation as it can be traced back to 1980s when the first metropolitan computer network was created. Since then the national strategy has been developed to improve ICT infrastructure. One of the strategies for ICT education, Cross Sector Strategy is to equip all schools with computer labs with 25 students for 1 computer in 2010 (Albanian Government document, 2009). Year 2006 Albanian government received loan from the Word Bank to develop the Education Excellence and Equity Project. A major part of the initiatives is to improve the quality of learning and teaching conditions for all the students and teachers of primary and secondary schools. With the relevant ICT available, establishing a digital library becomes critical as currently the amount of printed bibliographic resources available to students and teachers of Albanian primary and secondary schools is very limited and there are almost no learning resources

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 341–347, 2011.

of digital formats to the same user groups. To ensure the success of developing the digital library, the World Bank assisted Albanian government in executing the project by publically recruiting an international consultant whose responsibilities include but not limited to identifying the key stakeholders, defining Terms of Reference, developing implementation plan, providing budget projection, and planning intermediate execution.

2 Challenges and Significance of Developing the DL

Although ICT is a key element to be considered in building a digital library, accessing to ICT itself is not enough. Several major factors that affect the development of digital libraries are cost, resource availability and accessibility, efficiency of Technology, and impact of social context. As many of other developing countries, Albania is lagging behind in both print and digital resource development. Primary and secondary schools of Albania are particularly lacking of resources of any format due to shortage of funding to access or purchase resources. As major part of the strategy of developing eGovernment, and eSchools, Albanian government set top priority for improving education system. Commitments are being made to address digital divide issues by providing Internet connectivity and ICT and improve learning environment by building a digital library for pre-university schools. The intended DL will be the first digital repository nation wide. Therefore, the success of the project will have overarching impact on future digital projects of such kinds. The co-author of this paper, Ardiana Sula, Head of Educational Technologies of Ministry of Education and Science briefed me, the selected consultant, with the current condition of online resource development of the country: the only online digital resource available to the public is the national library's OPAC (Online Public Access Catalog), which is in Albanian Language. There is no professional librarian who has had formal library school education in Albania. Nobody has knowledge of metadata or experience of managing digital materials. The government will identify a qualified third party company to develop the digital library with the assistance of the international consultant.

In order to better understand the IT infrastructure, bibliographic resource background, users' information needs, and ICT capability to deliver the digital information to the target users, I visited some primary and secondary schools that typically represent geographical differences: one high school and one elementary school in the capital city of Albania, Tirana, two high schools in a suburban area, and one elementary school and one junior high in a village about 70 miles away from Tirana. Each visit included talking to students, interviewing teachers and staff, and visiting the schools computer lab and library. The interviews were conducted with selected subject teachers, the administrative staff who are in charge of the computer lab or courses, and the school librarian. Here is a brief summary of the findings:

2.1 ICT Capability

All the schools are equipped with a computer lab. The number of computers varies depending on how many students have registered in the school. According to Cross Sector Strategy defined by Albanian government in 2007, the goal of PC distribution is 25 students to 1 computer in 2010, but the goal has not been reached. All these computer labs have Internet access through satellite (VSAT), which reflects the government's effort of promoting ICT usage in classrooms. The speed of Internet connectivity is about 512/128Kbps shared to no more than 5 users, and information in multimedia format such as audio/video is discouraged to play because of the networking speed limitations. Every school has one projector. The teacher needs to make a request when he/she plays computer files or CD/DVD in class. The time of using the device is limited to a class period. Computer lab is used when, most of the time, the teacher assigns homework that are required to use Microsoft Office applications to create and edit the assignment. According to the school principals some of students in Tirana have computers at home with Internet connectivity, but most of students in rural or villages do not have personal computers.

2.2 Information Needs and Information Availability

100 percent of these teachers said they need extra teaching materials or supplementary resources to their classes: maps, history materials of other countries, math exercises, physics and chemistry lab exercises, tests and language resources, only name a few. Currently what they have are exclusively textbooks. They have rarely used reference materials, and they have zero knowledge of scholarly databases or know little about electronic journals. Each school has its own library. The library collection consists of a small number of books, and most of these books are fictions. Funding for purchasing books is not always available. There is no catalog of any type that users can search; books are shelved and organized in a rather arbitrary way that the librarian can easily locate. The librarian manually records check-in and checkout activities. The materials in the library are not necessarily intended for the class teaching or learning. When teachers need more learning materials or assign projects to students that require additional information, they will search Google to get the information. However, there are often times their searches are not very successful because they do not know what an effective search strategy is.

2.3 Level of Information Literacy

Most of teachers admit they cannot tell the difference between Google and a digital library. When a teacher needs additional resources to assist the class, or he/she assigns students project to search online, Google is the main source that both the teacher and students use. They do not question the legitimacy of obtained information or evaluate the source of information since it is hard for them to distinguish opinions and facts. When they think the information from

the search results is what they need, they simply copy and integrate as part of their own work. They were not aware that they were risking of plagiarizing other people's work.

2.4 English Language Proficiency

Because most of high-quality resources are English-based information, it becomes a big obstacle to the teachers because most of the teachers do not possess basic knowledge of English Language. Although the useful information is available online, the language barrier prevents them from using the information. In addition, because Albanian Language is much less popular than those such as Spanish, French, Germany or Chinese, even though lots of online resources are available in multiple languages, Albanian may not be on the list of offered languages. Interestingly, lots of students have better English language skill due to the fact that they have language classes and they have learned a lot from watching English-language movies or TV shows.

After these visits, the director of ICT education said basically most of the public schools are in the same situation in terms of ICT facilities and available resources. The interviews clearly presented the challenges: while building a digital library that provides users needed information is essential, developing an effective digital library model that serves as a cornerstone for the nations future digital library development is even more important.

3 Designing Considerations

Studies from researchers of different disciplines have shown that a key element of measuring the success of a digital library is largely based on the level of user satisfaction. A user-centered information system will help to maximize the usability of the digital library by providing positive user experience, and thereby, understanding users' cognitive effort, and confidence, and searching behavior will be critical to design such system. Given the fact that the identified key stakeholders of DL are pre-university teachers and students who have low exposure to computers and Internet, and possibly high level of information illiteracy, these aspects have to be considered regarding designing DL:

- ☐ Information needs vs. the purpose of using the information
- ☐ Information availability vs. information diversity
- ☐ Search behaviors vs. age, and educational background
- ☐ Perceptions to the digital contents vs. information presentation, and the nature of information

Teachers and students play different roles in any educational systems, and therefore, their information needs and the purpose of using information may vary. With limited funding in this project, the coverage of resources will doubtlessly

be restricted. As a result, the availability of relevant content in local languages that are particularly useful to the target groups should be the top priority. Nonetheless, such priority should not narrow the scope and diversity of information availability. Although electronic resources are scarce in Albania, and acquiring electronic resources from other countries is expensive, there are possible alternatives that are worth exploring. Apart from prioritizing the funding distribution, free online resources including free digital libraries; electronic textbooks; open journals; Open Archive Initiative, etc. are effective way to expand the variety of digital collections, and lower the cost of obtaining resources. In addition, as many students have good skill of English Language, they will more likely to benefit from rich resources on the web that they were not aware.

However, resource adequacy does not necessarily result in satisfied users. A poorly designed information system due to "inappropriate implementation strategies of accessibility " (Adams, 2002) can significantly reduce the usability of the resources. Studies have found that the way a system is designed has big impact on Human-Computer-Interaction (HCI). "the properties attributed to the system as the interface, the language, the orientation on the tools and devices, the work load, flexibility, compatibility with other systems, communication, as well as the effort to work, intervene directly in this interaction" (Norman, 1986). A well-designed information system always reflects good studies of users searching behavior, background and experience. The target users in this project have limited experience in interacting with computers and searching Internet. Their extensive searching experience comes from using Google. For that reason, an interface that can help these users quickly get to the access point will surely help ease the frustration and create positive searching experience. Make the user interface transparent to your audience. It shouldnt be noticed. It shouldnt be something people have to learn or decipher. If everything is clean and clear, then learners can get on with the task of learning and not have to think about how to interact (Neilson, 2011). Besides an intuitive interface, efficient searching algorithm with ranking mechanism and full-text capability will certainly increase the resource accessibility. Because most of users do not have much knowledge of how information is organized, highly relevant searching results will encourage user's interaction with the system and arouse their interest in using the available resources.

It is arguably that maximized accessibility and efficient system can be more optimized by good ICT infrastructure and knowledgeable users. Albanian government has been making all possible efforts to improve ICT infrastructure, promote ICT usage in classrooms, and improve the quality of education, but the improvement takes time and needs strong financial support. Although all the schools have Internet access, the low speed of Internet connectivity prevents rich contents such as videos or audios from being used for teaching and learning purposes. Plus these users possess low level of information literacy as demonstrated in the interviews. They do not evaluate the information found on the World Wide Web, and seldom question whether the resource comes from personal site or reputable journal. For those reasons, the content selection has to

be scrutinized to both fit the current ICT capability and enable users to meet their educational purposes.

4 Opportunities of Bridging the Gap

Although funding is limited for this project, and building a digital library is very costly, it is not impossible to build the digital library economically and technologically suitable to Albania. Besides taking advantage of freely available electronic resources online, free open source software such as Greenstone, DSpace, Fedora, and CDSware etc. used to build the digital library system are available. These applications can alleviate the burdens of technological expenses and human talent restrictions, and make the development affordable. Since the beginning of 21st century, the government has made further effort to ameliorate ICT environment and educate the citizens. A crucial part of the government program is to realize computerization of all primary and secondary schools in the country. The e-Schools Project aimed to equip all 2,100 public primary and secondary schools with modern computer labs with Internet access, establish new ICT curriculum in all public schools and train teachers in ICT literacy has been accomplished; ICT-aided learning methodologies that create interactive school environment have been introduced. These projects are good steps towards improving information accessibility and bridging the digital divide.

5 Conclusion

Albanian primary and secondary schools, as those in other developing countries, are challenged by lacking of both print and electronic bibliographic resources. With the available funding from the World Bank, the goals of improving the learning environment and promoting ICT usage in classroom necessitate the cause of developing a national level DL that will elaborate the quality of teaching and learning pedagogically. The limited funding, overwhelming information need, poor ICT infrastructures and information illiteracy have created major challenges, but through careful considerations of designing and by taking advantages of existing electronic resources and open source software, Albania government will be able to build DL economically and technologically appropriate to the country, and the DL will provide opportunities for the country to bridge the digital divide.

References

1. Adams, A., Blandford, A.: Digital libraries in academia: Challenges and changes. In: Lim, E.-p., Foo, S.S.-B., Khoo, C., Chen, H., Fox, E., Urs, S.R., Costantino, T. (eds.) ICADL 2002. LNCS, vol. 2555, pp. 392–403. Springer, Heidelberg (2002)
2. Chan, L., Kirsop, B., Costa, S., Arunachalum, S.: Improving access to research literature in developing countries: Challenges and opportunities provided by Open Access. In: IFLA General Conference and Council, Oslo, Norway, August 14-18 (2005)

3. Clark, M., Gomez, R.: Cost and Other Barriers to Public Access Computing in Developing Countries. In: 2011 iConference (2011)
4. Ferreira, S.M., Pithan, D.N.: Usability of Digital Libraries: a Study Based on the Areas of Information Science and Human-Computer Interaction. In: World Library and Information Congress: 71st IFLA General Conference and Council, Oslo, Norway (2005)
5. Kuny, T., Cleveland, G.: The Digital Library: Myths and Challenges. IFLA Journal 24, 107–113 (1998)
6. Lalmas, M., Bhat, R., Frank, M., Frohlich, D., Jones, M., Dhwani, N.: Bridging the Digital Divide: Understanding information access practices in an Indian village community. In: 27th International Conference, SIGIR (2007)
7. Norman, D.A., Draper, S.W.: User-Centered-System Design: New Perspective on Human-Computer-Interaction. Lawrence Erlbaum Association, New Jersey (1986)
8. Spasser, M.: Realist Activity Theory for Digital Library Evaluation: Conceptual Framework and Case Study. In: CSCW, vol. 11, pp. 81–110 (2002)
9. Support to the Ministry of Education and Science for the Implementation of the e-Schools Programme. e-Schools Final Report. Albania (2009)
10. User Interface Design For eLearning, http://theelearningcoach.com
11. Witten, I.H., Loots, M., Trujillo, Bainbridge, D.: The Promise of Digital Libraries in Developing Countries. The Electronic Library 20, 7–13 (2002)
12. Witten, I.H., Loots, M., Trujillo, Bainbridge, D.: The Promise of Digital Libraries in Developing Countries. The Electronic Library 20, 7–13 (2002)
13. Witten, I.H., Bainbridge, D.: A Retrospective Look at Greenstone: Lessons from the First Decade. In: JCDL 2007: Proceedings of the 2007 Conference on Digital Libraries, New York, NY, USA, pp. 147–156 (2007)
14. Witten, I.H., Bainbridge, D., Boddie, S.J.: Power to the People: End-User Building of Digital Library Collections. In: Proceedings of the First Joint Conference on Digital Libraries, Roanoke, VA (2001)

DAR: Institutional Repository Integration in Action

Youssef Mikhail[1], Noha Adly[1,2], and Magdy Nagi[1,2]

[1] Bibliotheca Alexandrina, El Shatby 21526,
Alexandria, Egypt
{youssef.mikhail,noha.adly,magdy.nagi}@bibalex.org
[2] Computer and Systems Engineering Department, Alexandria University,
Alexandria, Egypt

Abstract. The Digital Assets Repository (DAR) is a system developed at the Bibliotheca Alexandrina to manage the full lifecycle of a digital asset: its creation and ingestion, its metadata management, storage and archival in addition to the necessary mechanisms for publishing and dissemination. In its third release, the system architecture has been revamped into a modular design including components that are best of the breed, in addition to defining a flexible content model for digital objects based on current standards and a focus on integrating DAR with different sources and applications. The goal of this paper is to demonstrate the building blocks of DAR as an example of a modern repository, in addition to discussing the challenges that face an institution in consolidating its assets and DAR's answer to these challenges.

Keywords: Institutional Repository, Integration, Modular Architecture, Digital Assets Repository, DAR.

1 Introduction

Given the explosion of digital content currently available and its variety, *Institutional Repositories* have become a vital and crucial component for any organization to preserve and manage the lifecycle of digital objects. Institutional repositories come in different varieties, however most of the current solutions are monolithic without enough flexibility to adapt their components as needs arise or they become too cumbersome to implement due to their complexity. Current solutions leave a lot to be desired in integrating applications that publish digital assets from the repositories in addition to integrating the repositories with different input sources. Bibliotheca Alexandrina (BA) developed its Digital Assets Repository (DAR)[1] to address these particular needs among other challenges that face repository administrators. DAR is an eco-system of components that manages the full lifecycle of a digital asset: its creation and ingestion, its metadata management, storage and archival in addition to the necessary mechanisms for publishing and dissemination. The design of DAR incorporates a modular best of the breed approach in different aspects of a modern repository. An API is provided that keeps applications in synch with the repository. Applications get the latest version of their digital objects or their metadata automatically once they are changed or added inside the repository. Several interfaces can be built on top of this API to integrate DAR with other systems thus extending its features.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 348–359, 2011.

DAR plug-in architecture provides flexibility to integrate with different sources of metadata, such as an ILS, other repositories or databases. This allows for collaboration with other repositories easily by ingesting some of their collections and synchronizing their metadata through plug-ins. DAR provides a flexible model to represent different types of digital objects. It also uses well established standards like METS[2] and MODS[3] for metadata. Usually there is a need to digitize an item before the metadata is ready. DAR allows the ingestion of the object in an intermediate state. Once the metadata is ready, the item becomes qualified for dissemination. After ingestion, the metadata of the object is kept in synch with the metadata sources, and can be updated by human operators if no metadata source exist.

Institutions also face a daunting problem of having several applications as separate silos where each application hosts a copy of the objects. This causes redundancy, divergence and failure to manage the original objects in a global consistent manner. DAR addresses this by managing one instance of the object inside the repository. DAR uniquely identifies objects using a persistent handle. Objects can be grouped into sets, and a single object can be a member of different sets. This allows administrators to share the objects among several applications where each application has access to particular sets of objects. Different applications can maintain different derivatives of this same object independently. DAR heavily relies on RDF relations to define sets and relations between objects.

This paper is structured as follows: Section 2 presents some of the related work. Section 3 gives an overview of the system architecture and sections 4 through 8 describe in detail the main system components. A scenario for integration in included in Section 9. Section 10 concludes and presents directions for future work.

2 Related Work

Due to the increasing need to preserve and manage the large amount of digital assets currently owned by institutions, several repository solutions have emerged. EPrints [4] provides a solution to preserve scientific data, theses, reports, and multimedia, with optimized support for Google Scholar focusing on open access research. Greenstone [5] is an open source software solution for building and distributing digital library collections. DSpace [6] is commonly used by academic and research libraries as an open access university repository for managing faculty and student output and has the largest installation base. DSpace offers a full application and not just a framework with built in full text search based on Lucene [7]. It uses qualified Dublin Core as the default metadata schema. The previous alternatives aim at providing an out-of-the-box complete experience and ease of adoption. However, detailed configurations and customizations require more effort and time and extensibility is limited. They do not offer the level of metadata representation flexibility, scalability, or modularity required for large repositories with very specific requirements. Fedora [8] is a conceptual framework that uses a set of abstractions for expressing digital objects providing the basis for developing software systems for searching and administration through the provided web APIs. In Fedora, content is managed as data objects, composed of components ("*Datastreams*") that contain either the content or metadata about it. It relies on a Content Model Architecture [9] to represent objects. Fedora does not provide

an out-of-the-box solution, but rather requires programming expertise to setup the repository. Fedora's flexibility however makes it sometimes too cumbersome to implement in addition to requiring a certain level of programming expertise to build on top. DuraSpace [10] announced that DSpace and Fedora will be merging into DSpace with Fedora inside in 2011-2012, to retain the out-of-the-box experience that is DSpace, while also enabling the extra features that Fedora provides, e.g. versioning, relationships between objects, flexible architecture [11].

Stemming from the fact that most of the current solutions are monolithic without enough flexibility to adapt their components as needs arise or they become too cumbersome to implement due to their complexity, the search for the best solution for adoption has been the focus of many entities. The Arrow project [12] identified and tested solutions to support best practice institutional digital repositories for universities in Australia. It found that no one system or workflow will suit every university. Arrow recommended Fedora as a repository platform layer managing the object access and metadata combined with VITAL [13] as a services layer on top providing workflow extensions and advanced searching capabilities. The Hydra Project [14], a multi-institutional collaboration effort, on the other hand aims at providing an open source framework featuring the Fedora Repository on the back end, with a front end comprising Ruby on Rails, Blacklight [15], Solr [16], and a suite of web services providing similar functions. eSciDoc [17] also uses Fedora in its infrastructure to manage the digital objects while providing a set of loosely coupled open source services in a service-oriented architecture on top that can be used in building applications. The Stanford Digital Repository SDR [18] adopts Fedora in its Digital Assets Registry and as a metadata management system. *Digital stacks* comprise a suite of Ruby on Rails-based applications, with Solr index metadata stores providing asset discovery and delivery. SDR relies on Tivoli Storage Manager from IBM as its storage management subsystem. The SPAR project [19] at the BNF considers storing the Metadata using RDF in the Virtuoso triple store [20] the least risky approach [21] while still accepting a SIP with a METS manifest. It relies on iRods[22] for the storage subsystem.

DAR combines the best of the breed technologies to provide a modular repository based on latest standards with tools to manage digitization and encourage metadata entry by normal users. DAR data model is capable of representing all types of digital material. It focuses on the integration with different sources of metadata and objects through its flexible plug-in architecture, in addition to providing an API to integrate with applications that publish the digital objects stored within the repositories.

3 System Architecture

As shown in figure 1, DAR is divided into four main components: The *Digital Assets Factory* (DAF) provides a unified means of ingestion into the system from multiple sources through its ingestions plug-ins. A *Digital Assets Metadata* (DAM) subsystem manages the metadata even in an incomplete state. *Digital Assets Publishing* (DAP) components allow applications to synchronize objects and their metadata stored in their databases/indexes with the repository. Finally, the *Digital Assets Keeper* (DAK) manages access to the object files, versions and caching.

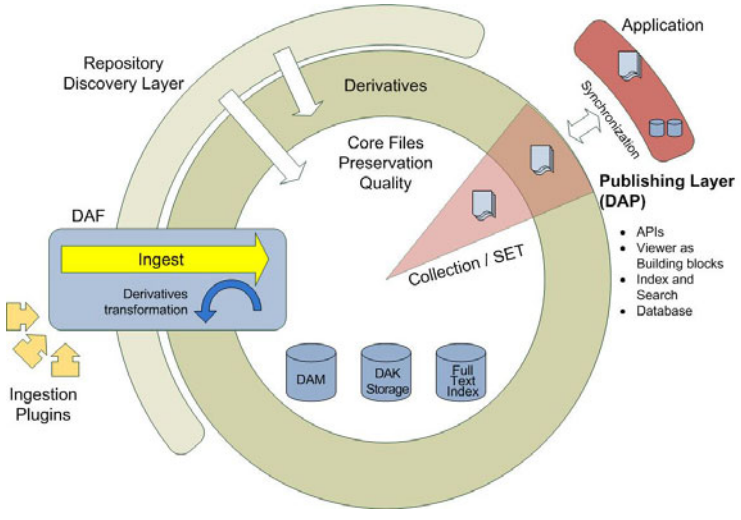


Fig. 1. DAR Conceptual overview

To facilitate object re-use, objects inside DAR are consolidated into sets or collections, and objects can belong to different sets. The core files, in preservation quality, are kept online on spinning drives for access by publishing applications. DAR also stores simple derivatives for all objects used for display by the *Discovery Layer*. Through the discovery layer, users of the repository can browse and search all assets stored within using simple viewers. A full text index based on Solr[22] search engine provides morphological search across the metadata and textual content of the objects stored in the repository.

Figure 2 depicts the detailed architecture of DAR in its latest version. In designing DAR, BA decided to rely on a modular design that allows it to mix the best of the breed components. The system incorporates components like Fedora, Mulgara[23], the Handle[24] system and search engines like Solr to manage metadata in a synergetic way. This modularity is possible through abstractions on the component level. A *RESTfull* API wraps the repository exposing its features to authenticated users. The API is used for ingestion of objects, metadata synchronization, discovery of items from the discovery layer and for application integration.

DAR uses METS to define objects stored within the repository. MODS is used for descriptive metadata for most of the object types. DAR uses Fedora as a *Metadata Registry* to manage the metadata. More information on that is provided in Section 5. Defining relations between objects is also important. RDF relations between objects are stored in Mulgara *Triple Store* providing facilities to run queries about set membership and object relations. Each object inside the repository is uniquely identified using a UUID. The UUID is used to generate a persistent Handle for that object making it possible to refer to the object consistently. A list of external identifiers are also stored with the object to provide references to the source of the object or to store any other form of identifiers that is specific to this type of object, e.g. ISBN for books.

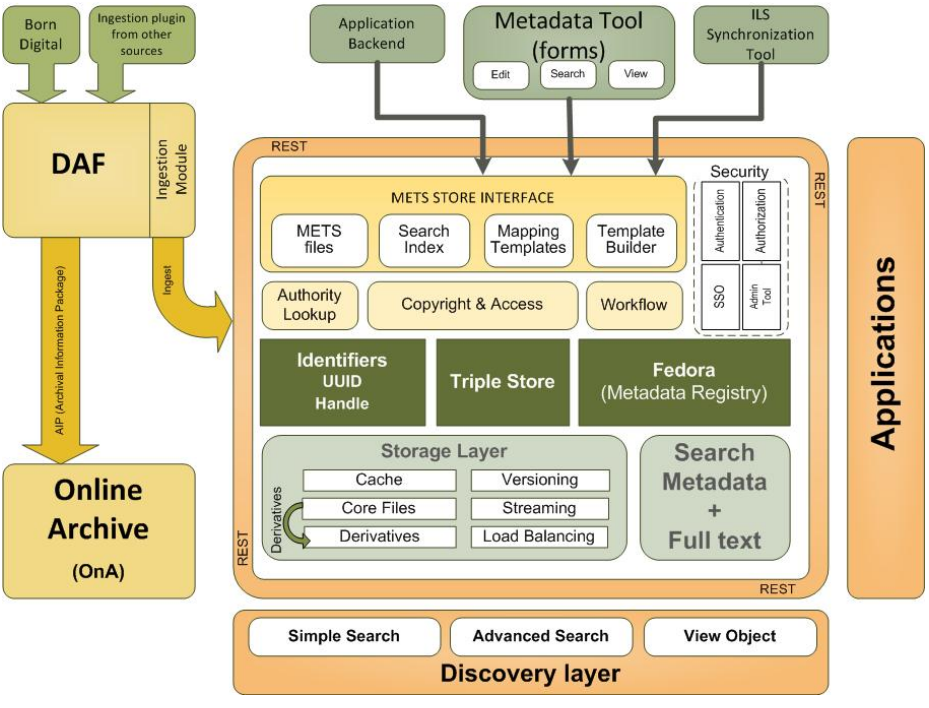


Fig. 2. DAR detailed architecture

A *Security* subsystem provides authentication and access control policy enforcement for specific sets or objects. DAR provides LDAP integration in addition to a local user database. A flexible authorization model enables administrators to define the different access levels for certain objects or sets. A *Single Sign On* module allows repository users to login once and gain access to all components of the system. A *Storage* subsystem provides a storage abstraction layer to isolate the underlying physical storage of the objects, in addition to other services like caching, load balancing among storage nodes, versioning and derivative management.

4 The Digital Assets Factory (DAF)

Many institutions are currently digitizing their collections to facilitate their access to users. Digital collections are very diverse and contain different types of objects. Since there is no one-size-fits-all tool that can perform all the digitization tasks, each object type might rely on a very different tool set and process to perform the digitization.

The Digital Assets Factory (DAF) [25] provides a configurable and flexible management tool for any digitization workflow, integrating with the current tools used for digitization. A digitization workflow is defined as a set of phases (figure 3), e.g. scanning, processing, quality assurance, encoding …etc. Administrators can define the sequence of phases required for a digitization in addition to adding pre-phase and post-phase checks, making sure that the process adheres to the digitization standards

in the institution. DAF checks that the correct file types, number of files and naming conventions are in place before and after the current phase, and can manage several types of workflows for different object types. It integrates with automated phases offloading human operators to do tasks that humans are good at: e.g. OCR correction.

Understanding the importance of associating the object with its metadata as early as possible, DAF integrates with external sources of metadata, through the development of plug-ins, where the digital objects to be ingested have their metadata in an external ILS, repository or a database.

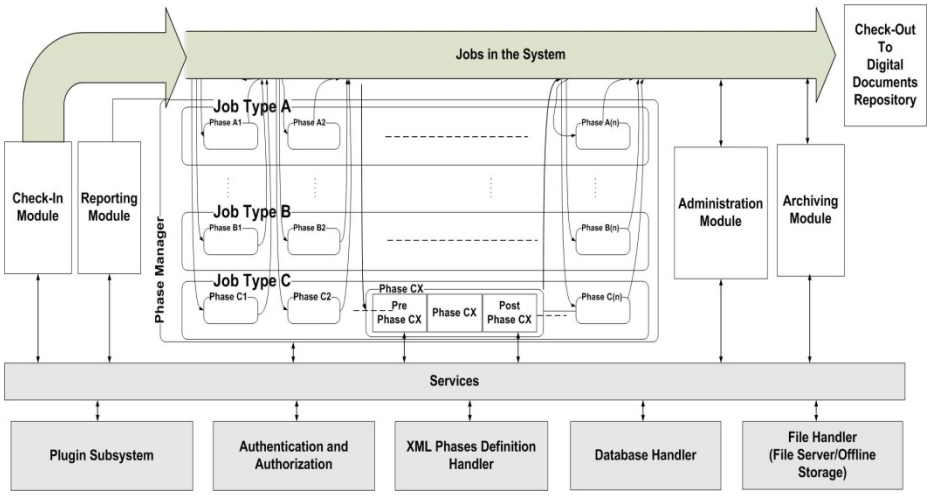


Fig. 3. DAF Architecture

Figure 3 demonstrates the system architecture of DAF: A *Check-in Module* adds jobs to the system. Operators will be able to work on these jobs through the various phases of the workflow managed by the *Phase Manager*. Each phase has a pre-phase and a post-phase check to verify the output. After the jobs complete the workflow, they are ingested into the repository through the *Check-out Module*, and an archival copy is sent automatically to the archive through the *Archiving Module*. If a job needs to be retrieved from the archive, DAF is used to fetch it and reload it into the system to perform the necessary actions upon it. DAF allows operators to request a re-do for a previous phase if they discover any problem. Operators can exchange messages indicating problems they face for a particular job so that other operators along the flow would take the necessary actions accordingly. The digitization supervisor checks the request, and can direct the job to another phase accordingly providing a means to correct the job as soon as the problem is detected. The quality of work and the performance of the digitization operators or automated phases can be monitored through the *Reporting Module*. DAF provides timely reports to various levels of management describing the workflow on a daily, weekly or longer basis allowing online queries about the current status of a certain asset during the digitization workflow. Reports display the number of jobs assigned to different phases of the workflow indicating the pending, running and finished tasks.

DAR is designed to be compliant with the OAIS [26] architecture since long term archiving of digital assets is crucial. DAF provides a unified means of ingestion into the system: in the pre-ingest stage, DAF handles both digital born objects and the digitization of physical objects. It creates the necessary SIP to be ingested into the repository, namely into the Digital Assets Keeper (DAK). DAF also creates an AIP that goes into the Online Archive (OnA) subsystem for long term archiving. Intermediate files created during the digitization are also included in the AIP for future reference. The item will be kept in sync with the metadata source and the AIP will be updated if any change occurs. BA provides DAF to the community as an open source tool (http://wiki.bibalex.org/DAFWiki).

5 The Digital Assets Metadata (DAM)

The wide spread and ease of use of technology has resulted in a plethora of different types of digital objects that an institutional repository should handle. Currently DAR holds more than 430,000 objects including books, photos, manuscripts, maps and documents with support for other types like videos, audio and more that are planned for ingestion. It is crucial to maintain the metadata along with the digital object. The object metadata are stored in Fedora, which is used as a metadata registry utilizing its different facilities to access the metadata.

5.1 The Content Model

DAR relies on well established standards to represent the metadata of the object, namely METS and MODS. However, it is a challenge to derive a model flexible enough to describe all types of library assets including books, maps, slides, posters, videos and sound recordings. DAR uses a hybrid *atomistic* and *compound* content model for objects providing flexibility of representation. An *atomistic* model represents every digitized piece of the object as a separate object, e.g. every photo of an album can be represented as a single object. Atomistic representation facilitates object re-use and discovery of a single piece of the object, e.g. a photo can be a member of several albums and can appear in a search result independently. Several objects can be aggregated into an *aggregate* object. An album containing multiple photos is an example of an aggregate object. The aggregate object holds metadata which is common across its child objects, e.g. album level metadata. When there is a need to create an album for photos taken in a certain conference grouped by conference sessions, photos of a certain session will be grouped into a session aggregate object, then session objects will be grouped into a conference aggregate object representing the conference event as a whole. There is no limit on the number of levels in this hierarchy.

The *compound* model represents the object as one single object containing the definition of the different digitized pieces of that object. Compound models greatly reduce the number of objects in the repository and thus the number of relations to be stored in the triple store making the maintenance of the object easier. DAR represents books as a compound object. Once the book is inserted into the metadata registry, it will contain a single data stream containing references to all pages, a convenient representation for a book since we will not be re-using its pages in another object. A

single page would thus not be considered as a separate object. Figure 4 depicts the content model used for representing a book. A bibliographic record translates into a parent aggregate object to reduce metadata redundancy across multiple volumes. A volume (Book Y) is represented as an object bearing the volume level metadata. A compound object (Book Y Content) holds the actual content. It is worth noting that several books can be grouped into sets (Set X). A book can be a member of different sets. DAR provides the necessary tools to manage the set membership.

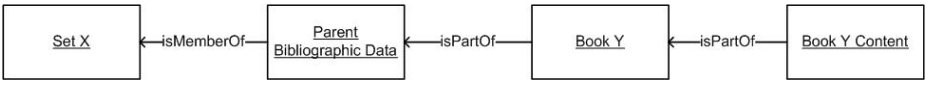


Fig. 4. A content model for books

The Fedora community defines a *compound* data model[9] as the representation used to define data objects containing different digitized pieces as multiple content-bearing data streams in a single object, e.g. a book containing multiple pages each represented as a separate data stream, or a book containing a PDF, DjVu, XML data streams. As described earlier DAR uses a slightly modified version of the compound model where there is a single content data stream containing the definition of the different pieces of the object. Gaining access to the pieces of the object is handled by DAR and not by the Fedora registry, thus making DAR independent from the Fedora-specific data stream representation of objects. This representation allows DAR more flexibility in handling the objects. DAR uses the same definitions for the *Atomistic* model as the fedora community.

5.2 The METS Store

The *METS Store* is a crucial component of DAM. When an object is ingested into DAR, a METS skeleton is first created storing the metadata provided upon ingest based on an XML template. The *METS Store* helps in handling one of the challenges faced by institutions, which is the need to quickly digitize some items for preservation where items are available for a limited time at the digitization facility with no or minimal metadata. The *METS Store* creates an intermediate incomplete state of metadata allowing the object to be ingested but the object is not yet ready for consumption.

Once the metadata is completed, either through synchronization with external metadata sources like an ILS or by a human operator, it is ingested into Fedora and the object is ready for usage through the publishing APIs. A simple yet flexible workflow engine handles these stages. Using this approach, a complete set of the metadata is kept in the *METS Store*, which acts as a backup for Fedora. DAR's modular architecture thus allows it to be independent of specific technologies. The metadata can be extracted from the *METS Store* and ingested into another system, and can also be used to reconstruct Fedora if any failure occurs.

5.3 Metadata Synchronization

The synchronization of the object metadata with external sources is based on XML templates to allow for flexible integration. An XML template, based on XSLT, is

created to translate the output of the ILS or a database into the necessary MODS representation. This translation is handled by the *METS Store*. In case there are no sources of information to extract the metadata, human operators, assisted by authority lists, can complete the metadata through the use of dynamic forms in the *Metadata tool*. The tool generates human friendly metadata forms through configurable XML templates that suit the needs of the users. This tool is intended for use by normal users and not necessarily experienced librarians or catalogers. It displays a reduced version of the digital object to assist the user in filling the metadata. The XML templates define the necessary input fields, a friendly name for the fields in a particular project, input type validations for different data types and predefined list of values. The template also defines tool tips to assist the user in understanding what to be written in each field. The *Metadata tool* represents the objects in a tree hierarchy depicting the sets and objects within. When several objects share some metadata, the *Metadata tool* provides a facility to copy the metadata to objects in lower levels in the tree. It warns the users if conflicts or overwrites would occur.

The *Metadata tool* allows the operators to work on the object according to their tight schedules. An operator can chose to edit an object then save it for later completion. To ensure the quality of the metadata, the *Metadata tool* allows the designation of several roles: editors and reviewers. After the editor finishes his work, the object metadata is inspected and approved by a reviewer. Once approved, the metadata is considered ready for ingestion into Fedora. The workflow engine handles this review workflow. Users can browse their tasks whether pending editing, partially edited, pending review and tasks that have been approved. Full text morphological search based on Solr is provided to allow the operators to search across the metadata of the items in their collections or sets to retrieve an item for further editing.

5.4 The Copyright and Access Module

A flexible *Copyright and Access module* manages access to digital objects. Copyright information is stored with the object. An application provides the necessary authentication information to request access to its sets of objects in addition to the level of access required including viewing, downloading, printing, etc. The *Copyright and Access module* also coordinates the access of an object based on the number of licenses available. If the institution owns the right to display the object a certain number of times simultaneously, the module would coordinate the use of different applications to make sure the maximum number of licenses is not violated. This service is exposed through the REST API where applications attempt to obtain a license before displaying the objects. Based on the access policy associated with the object, a particular audience can be granted access to a partial view of the object, while a full view can be provided to another audience set.

6 The Digital Assets Keeper (DAK)

The Digital Assets Keeper is responsible for keeping a “working” copy of the object online used for consumption. A complete archival copy of the item is deployed into the *Online Archive* (OnA). DAK maintains a unique copy of the object and every ob-

ject inside the repository is assigned a *persistent identifier*. DAK manages different versions of items. We are currently investigating the usage of pair-tree [27] structures among other approaches to provide a scalable and a flexible representation for object versions. A *storage abstraction layer* is used to isolate the repository from the underlying storage implementation. DAR requests access to an object through the object's identifier and a resolver would do the rest. This allows the implementation of different storage policies and several tiers of storage based on the frequency of use and other factors: e.g. frequently used objects can be kept on the fastest storage tier. The Storage layer handles load balancing across storage nodes in the same tier in addition to handling the caching of derivates used for the repository discovery layer and streaming of media files stored within the repository.

7 The Digital Assets Publishing (DAP)

There is a need to have applications highly integrated with the repository rather than being separate silos. Usually applications take a copy of the items, then they lose synch with the repository: they add, delete and modify the original objects without referring to the repository. DAR provides an API that allows applications to become repository-bound in a sense that when objects, that are in sets or collections the applications have subscribed in, are added, deleted or modified in the repository, the applications are notified and can request the latest updates of the object or the metadata through a RESTful API. One consistent instance of the object is kept in the repository and applications can focus on providing rich interfaces, creating their own derivatives of the objects while maintaining a link to the original object through the API and receiving updates.

Several applications have been built using this approach. A *Discovery layer* is provided for internal use inside BA giving the users access to the items in their totality inside the repository through simple viewers. Specialized *viewers* have been built to display items stored within the repository, such as books and photos. More viewers are still under development to provide unified access to the objects across applications built on top of the repository: e.g. tiled image viewer and manuscript viewer. *DAR books* (http://dar.bibalex.org) is an application built on top of DAR that displays the books stored in the repository (185,000 books) in a user friendly manner, providing browsing by facets, full text search and many other features. Whenever a book is added to or updated in DAR, it is automatically retrieved by *DAR books*. Another example is the print on demand (POD) integration layer that makes part of the content of DAR available through the POD system. Several interfaces can also be built on top of this API to integrate DAR with other systems.

8 The Online Archive (OnA)

The *Online Archive* (OnA) is a complete hardware and software solution that provides an underlying reliable and scalable archival storage. OnA ensures that any AIP ingested is mirrored at least once to provide redundancy. It heavily relies of checksums to ensure the integrity of the files at different stages. Given the exponential increase in

the size of data to be archived, the OnA provides a low cost scalable storage system based on commodity hardware with spinning hard drives. It runs special software developed by BA for data management.

9 DAR Integration in Action

In the following section, we will introduce an example that describes the life cycle of a digital object in DAR. Suppose that a group of digital objects are part of a collection donated to the library by a certain organization X. The collection's metadata is availed through an OAI-PMH interface. The objects received still require further processing at the digitization facility at the library, e.g. image processing and OCR, then they are to be added to an already existing application Y that uses DAR's API.

A DAF plug-in is built to ingest the objects into the digitization workflow. Once the processing is done, the objects are archived and ingested into the repository in an intermediate state where a METS skeleton is built. Another Plug-in is developed to synchronize the metadata obtained through OAI-PHM with the *METS Store*. Once the synchronization is complete, the objects' metadata is ingested into Fedora, indexed and the archive is updated. The objects are now accessible through the API. Since these objects should be added as part of application Y, the administrator adjusts the set membership for these objects to application Y. The application therefore discovers the objects the next time it asks the API for updates and loads the objects to cache them on its servers. When the metadata of an object changes at organization X. The synchronization plug-in loads the new values. The *METS Store* and Fedora are updated and re-indexing of the object is triggered. Application Y detects that the metadata of the object is updated so it loads the updated values into its database and displays them.

10 Conclusions and Future Work

DAR is the flagship of Bibliotheca Alexandrina's digital library. We have presented in this paper DAR's architecture and the main philosophy behind its design. DAR addresses the main challenges faced by digital repositories including supporting different digital formats, digitization workflows, preservation of digital material and content dissemination.

At version 3.0, DAR's overall architecture and design are established. Most of its core components have been developed, using open source tools, and deployed with several applications launched on top. A complete data migration was completed in December 2010. Development is underway to enhance the Storage layer component and versioning, in addition to extending the *Copyright and Access module* and adding virtualization support. The potential of Linked Data, Semantic Web and triple stores is yet to be exploited further in DAR. Scalability issues related to the number of RDF relations are also been studied and other triple stores are currently being tested for scalability. BA is currently working on the migration of existing applications into the repository modifying them to be repository bound thus consolidating the digital assets.

References

1. Saleh, I., Adly, N., Nagi, M.H.: DAR: A Digital Assets Repository for Library Collections – An Extended Overview. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) ECDL 2005. LNCS, vol. 3652, pp. 116–127. Springer, Heidelberg (2005)
2. Metadata Encoding and Transmission Standard (METS), http://www.loc.gov/standards/mets/
3. Metadata Object Description Schema (MODS), http://www.loc.gov/standards/mods/
4. EPrints, http://www.eprints.org/software/
5. Greenstone, http://www.greenstone.org/
6. Dspace, http://www.dspace.org/
7. Apache Lucene, http://lucene.apache.org/java/docs/index.html
8. Fedora Commons, http://fedora-commons.org/
9. The Fedora Content Model Architecture (CMA), http://fedora-commons.org/documentation/3.0b1/userdocs/digitalobjects/cmda.html
10. DuraSpace, http://www.duraspace.org/
11. DSpace Fedora integration, https://wiki.duraspace.org/display/DSPACE/Fedora+Integration
12. The Arrow Project, http://arrow.edu.au/
13. VTLS VITAL software, http://www.vtls.com/products/vital
14. The Hydra Project, https://wiki.duraspace.org/display/hydra/The+Hydra+Project
15. Project Blacklight, http://projectblacklight.org/
16. Apache Solr, http://lucene.apache.org/solr/
17. eSciDoc, https://www.escidoc.org/
18. Cramer, T., Kott, K.: Designing and Implementing Second Generation Digital Preservation Services: A Scalable Model for the Stanford Digital Repository. D-Lib Magazine 16(9/10) (September 2010)
19. The SPAR Project, http://www.bnf.fr/en/professionals/preservation_spar.html
20. Virtuoso Universal Server software, http://virtuoso.openlinksw.com/
21. Fauduet, L., Peyrard, S.: A Data-First Preservation Strategy: Data Management In SPAR. In: iPres 2010 Vienna, Austria (September 2010)
22. iRODS, http://www.irods.org/
23. Mulgara Semantic Store, http://www.mulgara.org/
24. The Handle system, http://www.handle.net/
25. Yakout, M., Adly, N., Nagi, M.: Digitization Workflow Management System for Massive Digitization Projects. In: 2nd International Conference on Universal Digital Library ICUDL 2006, Alexandria, Egypt (November 2006)
26. Reference Model for an Open Archival Information System (OAIS), http://public.ccsds.org/publications/archive/650x0b1.PDF
27. Abrams, S., Kunze, J., Loy, D.: An emergent micro-services approach to digital curation in-frastructure. In: iPRES 2009, Mission Bay, San Francisco (October 2009)

Linking Archives Using Document Enrichment and Term Selection

Marc Bron, Bouke Huurnink, and Maarten de Rijke

ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam
m.m.bron@uva.nl, b.huurnink@uva.nl, derijke@uva.nl

Abstract. News, multimedia and cultural heritage archives are increasingly offering opportunities to create connections between their collections. We consider the task of linking archives: connecting an item in one archive to one or more items in other, often complementary archives. We focus on a specific instance of the task: linking items with a rich textual representation in a news archive to items with sparse annotations in a multimedia archive, where items should be linked if they describe the same or a related event. We find that the difference in textual richness of annotations presents a challenge and investigate two approaches: (i) to enrich sparsely annotated items with textually rich content; and (ii) to reduce rich news archive items using term selection. We demonstrate the positive impact of both approaches on linking to same events and linking to related events.

1 Introduction

News, multimedia, and cultural heritage archives are opening up and publishing their content online, enabling users to search for items of interest across multiple archives. With the general public gaining access to archive content, an increasing number of users can be expected to exhibit exploratory behavior [2], rather than directed search typical of professional users [10]. In order to make archives accessible to the general public, modes of access supporting exploratory behavior should be examined.

One way to enable exploration over (multiple) archives is to create links between individual items. On the web a common method to enable exploration is to create hyperlinks between documents allowing a user to wander from one document to the next and gradually explore a topic of interest. In an archival setting the creation of links has received little attention, likely due to the focus within archives on annotation and preservation of individual items, rather than on supporting browsing behavior.

We examine the linking problem in an archival setting, focusing on events. Here we aim to connect an item from one archive to items in another archive that discuss the same or related events. Links to items describing the same event allow users to access different views of the same event, while links to items describing related events allow users to explore interconnected relationships between events. Such event-based links are particularly valuable for exploring news archives. We focus on a specific instance of the task: linking items from a newspaper archive with a rich textual representation to items from a multimedia archive that tend to have sparse annotations, see Figure 1.

Our scenario is characterized by two somewhat complementary challenges. First, the targets of our linking task—archived multimedia items—are relatively sparsely annotated which leads to recall problems. Second, the source of a link is a news article in

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 360–371, 2011.

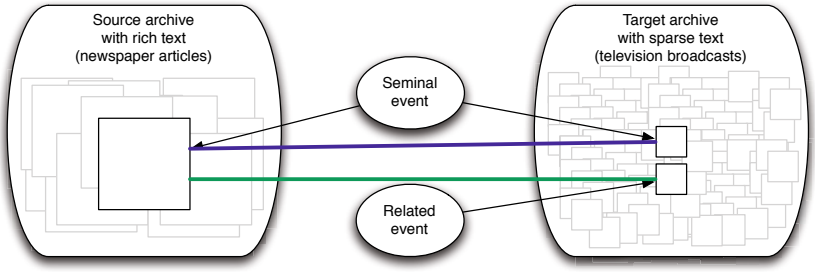


Fig. 1. Illustration of the event based linking task between archives

a news archive; such articles may be long and discuss issues that are only indirectly related to the seminal event that triggered the article, thereby potentially giving rise to a precision problem. These two problems motivate our primary research question: *when linking from an archive with rich textual content to an archive with sparsely annotated items, how can we compensate for the relative richness and sparsity of the representations to improve linking performance?* We divide this into two sub-questions:

RQ1. Does expanding sparse item representations with text from other sources improve linking performance?

RQ2. What effect does modeling reduced versions of the original richly represented source item have on linking performance?

We approach the linking task as a retrieval problem: given a source item, retrieve target items it should be linked to. To address our first research question we improve the representation of target items by enriching their sparse annotations with text from the target archive, the source archive, and Wikipedia. To address our second research question we reduce the representation of the source item by selecting a subset of terms from the source article, from manually created annotations or from automatically created ones.

The contributions of this paper are three-fold: (i) we define and motivate a new task, i.e., linking archives based on events, and identify future directions; (ii) we report on a set of experiments investigating the effect of item representation along two dimensions; and (iii) we demonstrate the effectiveness of linking based on enrichment of sparsely annotated items with content from a different archive.

We discuss related work in §2. Then in §3 we describe our item enrichment and linking approaches. In §4 we describe our data sets and experimental setup. In §5 we report our results and provide an analysis. We conclude in §6.

2 Related Work

We discuss work related to our task of event-based linking from a textually rich to a textually sparse archive. We consider links at the item level, i.e., linking from one item to another item, rather than linking phrases and words from an item representation such as is done for example in Wikipedia linking [15, 16] and hypertext linking [4].

One example of linking between items is the *alignment* task: identifying items in comparable collections that discuss the same person, entity, or concept. This task is

addressed by [11], who use a retrieval engine to align entries from four encyclopedias. They find that using document structure such as the title and body of an encyclopedia entry is an important component of achieving successful alignment. Our work differs in three respects: we link from an archive with textually rich items to an archive with sparsely annotated items; we include two types of linking; and we enrich sparsely annotated items.

Another area to cover the linking task is *topic tracking*, in which items are connected when they discuss the same seminal event or related events [1]. Commonly, this is done within a collection consisting of either a single news source [7] or a collection of multiple textual news services [17, 21]. Work on topic detection and tracking includes work on detecting novelty and redundancy using language models [21] and new event detection using an adaptation of the vector space model with named entities [12]. These methods use techniques from information retrieval to find link targets, based on similarity. We will also follow a retrieval-based approach, but as we are dealing with archives with disparate quantities of text, we focus on the effect of using document expansion and term selection techniques for linking.

The news context has also seen work into linking news articles to blog post entries that discuss them. Using the structure of news articles (title, lead, body, etc.) to model a query can help in linking to correct blog posts [20]. An early paper on the topic of cross-media linking investigates generating connections between news photos, videos, and text on the basis of dates and named entities present in texts associated with the items [3]. [14] investigated cross-media news content retrieval to provide complementary news information. This was done on the basis of news articles and closed captions from news broadcasts, and focused on differences in topic structure in the captions to find complementary news articles for broadcasts. Also relevant is work on linking passages from the closed captioning of television news broadcasts to online news articles [9]. Here, the focus was on the time-based aspect of identifying articles about the news subject being discussed at any particular point in time. An interesting finding was that term selection was valuable in identifying the correct relevant articles. We also apply term selection to our source items, but link to sparse content and additionally experiment with expanding target items.

The problem of text sparsity in the linking task has seldom been addressed. However, in the field of of information retrieval, it has been found that when faced with very short documents (i.e., documents with sparse text), *document expansion* can help improve retrieval performance [19]. Document expansion refers to combining text from related documents with the text of an original document. For a more in-depth study of expansion in the context of traditional document search, see [5].

3 Approach

We formally define the variants of the linking archives tasks addressed in this paper: *same event linking* and *related event linking*. Given an event e, described by a source item s from a source archive A_s with rich text representations, create links to target items $T = \{t_1, \ldots, t_n\}$ in a target archive A_t, where the event described by each $t_i \in T$ is the *same* as e. In the second variant, each $t_i \in T$ is *related* to e. The notions of same and related event will be defined in §4 below.

An item is *textually rich* when it contains, on top of human-annotated metadata, textual content. *Sparse* representations only contain human-annotated metadata. In the specific setting in which we are working, source items are news articles, hence textually rich, and target items are sparsely represented items in a video catalog (see below).

Linking model. We model the task of linking archives by finding a ranked list of target items t whose representation is most similar to the representation of the source item s. We use the vector space model [18] as our similarity function:

$$\text{sim}(s,t) = \frac{\boldsymbol{V}(s) \cdot \boldsymbol{V}(t)}{|\boldsymbol{V}(s)||\boldsymbol{V}(t)|}, \tag{1}$$

where $\boldsymbol{V}(s)$ and $\boldsymbol{V}(t)$ are vector representations of s and t, respectively, the numerator is the dot product of the vectors and $|\boldsymbol{V}|$ is the length of $\boldsymbol{V}$.

We compute the similarity between a fixed source item s and every potential target item t in the target archive and rank each t according to its similarity. The resulting ranked list is cut off at some rank n so as to yield a list of link targets.

Document expansion. To address sparseness of the representation of a target item t we use other items x for expanding the representation of t. Below we consider multiple sources A_x for the expansion items x: the source archive A_s, the target archive A_t or even an external archive A_e. To obtain expansion items x we compute the similarity between $t \in A_t$ and each item in expansion archive A_x and rank its items by similarity, as in (1). The resulting ranked list is cut off at some rank m to yield a list of expansion items; these are then concatenated to t to form an expanded representation of t.

Selecting representative terms. Recall that our source items are textually rich. To address the potential of topic drift that may result from textual richness, we investigate the effect of automatically selecting a small number of terms from the text associated with a source item (instead of using all terms) when ranking candidate target terms. To select terms, we take the top $k\%$ terms from s ranked by their TFIDF score, which is defined as:

$$TFIDF(a) = \frac{c(a,d)}{|d|} \cdot \log\left(\frac{|D|}{|\{d \in D : d \text{ contains } a\}|}\right),$$

where $c(a,d)$ gives the count of term a in document d, $|d|$ is the length of a document and $|D|$ is the number of documents in the collection. As is well-known, TFIDF assigns more importance to terms that have a high frequency in a few representations, rewarding terms that are discriminative for a specific representation.

Selecting representative entities. We refine the selection of representative terms by only considering named entities. Named entities are a special type of term found to be important in identifying related events [12]. To select entities we apply a named entity recognizer [6] based on conditional random fields to the content of all source archive items. We then select the top $k\%$ entities based on their TFIDF value as with the terms.

Date filter. Finally, we also examine the use of the date field present in the metadata of both source and target items. Not only is the date field one of the most consistently filled fields in archival data, but it has also been shown that dates are useful when detecting same events [13]. We use a simple date filter that only allows a link from a source item s to a target item t if t's date is within an N day window around the date of s.

4 Experimental Setup

In this section we describe our experimental setup. We start by describing the collection used for evaluating the linking rich-to-sparse archive task, and follow with a description of our experiments with document expansion and term selection.

4.1 Evaluation Collection

Our evaluation collection consists of a source archive containing textually rich newspaper articles and a target archive of textually sparse television news broadcasts to which we want to link. We single out a set of source items as our test cases for linking, and for each test source item, we have a set of relevance judgments indicating which items in the target archive refer to (i) the same seminal event, and (ii) related events.

Source archive. Our source archive consists of 346,559 newspaper articles published by a Dutch newspaper, the NRC Handelsblad,[1] from 3 Jan. 2005 to 8 Jun. 2010. Each article consists of the article text (article title and body) and a series of metadata fields created by archivists at the newspaper. These metadata fields comprise of *persons*, *locations*, *organizations*, *events* and *keywords* that are the subject of an article. Rather than exploiting the detailed specifics of NRC's archive, we combine all of the data from the metadata fields for an article together; we refer to this aggregated set of items as the *metadata* for s. We refer to the article text as the *content* of s. On average, source item content has 409 terms and metadata has 8 terms, for a total of 417 terms per item.

Target archive. Our target archive in this paper consists of 73,666 television news stories obtained from the Netherlands Institute for Sound and Vision, the Dutch national audiovisual broadcast archive.[2] We restrict the target archive to news stories broadcast during a period that encompassed the period of the source article collection, (1 Jan. 2005–20 Dec. 2010). We limited the target archive to news stories as other program categories, e.g., game shows and soap operas, are unlikely to yield suitable link targets for news articles. Each news story is manually described by professional archivists, with free-text *description* and *summary* fields and structured fields describing *persons*, *locations*, *keywords* and *other names* that are the subject of the news story. Once again, rather than considering the text of all these fields individually, we combine them to form the *metadata* for a given target item t. On average, target item metadata consists of 13 terms, illustrating the relative sparsity of text as compared to the source archive.

Events. We use the definition of event used at the Topic Detection and Tracking (TDT) campaign which makes a distinction between *seminal events*, i.e., high impact news events that generate follow-up events, and *related events* that are caused or predict the seminal event but are not seminal events by themselves.[3]

Test source items. In order to evaluate our linking approaches, we select a set of source items to use as test items to be linked. We use two requirements for our selection: the

[1] http://www.nrc.nl/
[2] http://instituut.beeldengeluid.nl/
[3] http://projects.ldc.upenn.edu/TDT5/Annotation/TDT2004V1.2.pdf

selected item should contain a clear seminal event (to facilitate judgments of system-generated links) and there should be at least one item in the target archive that covers the same or a related event. To satisfy the first requirement, we randomly select news events from Wikipedia listings of important events per month[4] and manually search the source archive to identify a newspaper item describing the event. To satisfy the second requirement, we search in the target archive to make sure that there is at least one television broadcast that describes the same or a related event. If so, the item is selected as a test source item. In total we selected 50 test source items, describing a range of events such as *16 May 2007: Nicolas Sarkozy is sworn in as the new president of France*; and *7 July 2009: A memorial service is held in the Staples Center in Los Angeles for the deceased pop icon Michael Jackson*.

Relevance judgments. We create relevance judgments using the pooling method adopted by TREC [8], the de-facto standard for creating relevance judgments for small to moderately sized test collections. We performed pooling on the basis of the sets of results produced by different linking systems. For each system and source item, the top 20 ranked documents were selected for inclusion in the pool. These results were then merged and duplicated documents were removed. The merged lists of results were then shown to human assessors, with results for each individual source item being judged by the same assessor to ensure consistency of results. The assessors were instructed to make a distinction between target items that describe the same event as the source item and targets that describe related events. The assessors' instructions were taken from the TDT assessor manual.[4] Due to the nature of related events, for each source item the average number of target items describing the same event is much lower than the number describing related events (2.4 vs. 11.8).

4.2 Experiments

Recall the two main research questions from § 1. (RQ1) Does expanding sparse item representations with text from other sources improve linking performance? (RQ2) What effect does term selection on the original textually rich source item have on linking performance? Below we list the experiments we conduct in order to answer these questions. All experiments are performed on the two tasks: *same event linking*, i.e., linking to items that describe the same event, and *related event linking*, i.e., linking to items that describe a related event.

Baseline. As our baseline we perform linking using all of the source item's content and metadata without term selection to representations of target items without expansion.

Expanding sparse text representations. In order to answer our first research question we evaluate the effect of increasing the number of documents used to expand target items on linking performance. An overview of the experiments is given in Table 1. We experiment with three sources of information for document expansion: the target archive itself, expanding target items with representations from other items in the archive; Wikipedia, the online encyclopedia; and the richly represented, news-focused items in the source archive.

[4] See e.g., http://nl.wikipedia.org/wiki/Januari_2009

Table 1. Description of the expansion models. In all cases the original sparse target metadata is concatenated with n expansion documents to form the expanded item representation.

Exp. model	A_x	Description
baseline	–	no expansion
n target docs	A_t	add $n \in \{1, \ldots, 10\}$ documents from target archive
n Wikipedia docs	A_e	add $n \in \{1, \ldots, 10\}$ documents from the Wikipedia encyclopedia
n source docs	A_s	add $n \in \{1, \ldots, 10\}$ documents from source archive

Table 2. Descriptions of the term selection models evaluated in the term selection experiments. In all experiments the number of terms in the *source* item representation (content and metadata) is reduced. Target items consist of their original metadata without expansion.

TS model	Description
baseline	all text associated with s, including content text and metadata text
content	s content text only
metadata	s metadata text only
title	s title
lead	first 2 sentences of content s
x% terms	select top x% terms from *content* s, using TFIDF ($x \in \{10, 20 \ldots, 100\}$)
y% ne	select top y% entities in *content* s, using TFIDF ($y \in \{10, 20 \ldots, 100\}$)
combined	combine metadata s with optimal x% *terms* and y% *ne* from content s

Term selection for rich text representations. In order to answer our second research question we evaluate the effect of reducing the amount of text in a source item on linking performance. An overview of our term selection experiments is given in Table 2. First, we experiment with using newspaper article structure to reduce the source item representation, following the framework presented in [20]. We then experiment with using only the most representative unique terms and named entities in the source item. We also investigate using only the manual annotations, i.e., metadata, of a source item. Finally, we experiment with using the optimal combination of these options.

Evaluation measures and significance testing. We use three evaluation metrics for evaluating linking performance. Mean Average Precision (MAP), the average of the Average Precision (AP) scores over all test items. This metric evaluates the number of correct link targets in a list (of length 100 in our case), where correct targets higher in the list are assigned more importance. Precision at rank five (P@5) only considers link targets in the top five. A perfect score of 1.0 indicates that all five targets at the top are correct. When less correct targets exist the maximum score will be lower. Mean Reciprocal Rank (MRR) is the average of the Reciprocal Rank (RR) for each source item. The RR is the inverse of the first correct answer and indicates at which rank of the list of target items the first correct target is found. We use a standard paired t-test to determine significant differences between results. We use $^{\triangle}$ or $^{\triangledown}$ ($^{\blacktriangle}$, $^{\blacktriangledown}$) to indicate

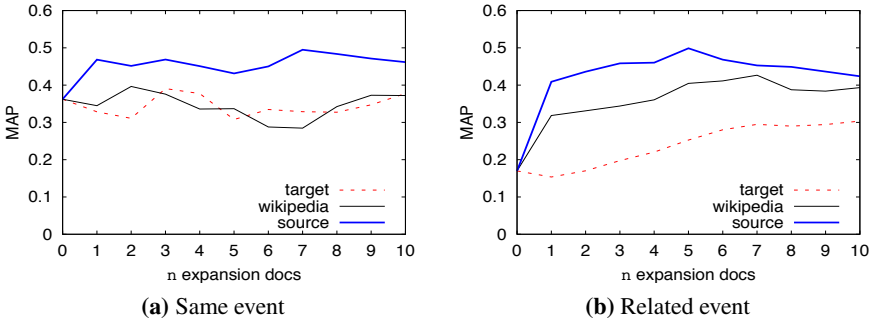
whether a score is significantly higher or lower than the baseline with a significance level of $\alpha < .05$ ($\alpha < .01$).

5 Results

5.1 Document Expansion

We first contrast different archives for expansion, i.e., the source archive, target archive, and Wikipedia. Figure 2a shows the MAP scores for *same event linking* using different expansion archives. Expanding with documents from archives other than the source archive does little to improve over the baseline even with the optimal number of expansion documents. Figure 2b shows that for *related event linking* expansion with documents from all three archives improves over the baseline. Again expanding with source archive documents achieves best performance. The performance scores for both event linking tasks, with the optimal number of expansion documents, are given in Table 3. The optimal number of documents to expand with from the source archive is seven for *same event linking* and five for *related event linking*; both yield a significant improvement over the baseline. We note that although optimized for MAP, the other early precision metrics follow the same trend in that the optimal number of documents for MAP is also the optimal number for the other metrics. The P5 scores for *same event linking* do improve (by 40.9%), but remain relatively low; this is due to the small number of relevant target items per test item (on average 2.4).

Let us examine the source item that benefits most from document expansion in the *same event linking* task. The title of the source item is "Openness expenses Dutch Royal Family." The description of the target item is: "Prime Minister Balkenende promises the House of Representatives transparency in the expenses of the Royal Family." The underlying event of the source and target item is the same, i.e., a parliamentary discussion about transparency with respect to the expenses of the Dutch royal house. However, the viewpoint of the event is described from a different angle in each item: the source item focuses on a request for more transparency from the house of representatives, while the target item focuses on the prime minister promising this transparency. Document



(a) Same event **(b)** Related event

Fig. 2. Document expansion with n documents, from the target archive, the Wikipedia encyclopedia, and the source archive. Here 0 indicates no expansion.

Table 3. Results of document expansion; significance is tested against the baseline

	Same event				Related event			
Exp. Model	*detail*	*MAP*	*P5*	*MRR*	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2000	.4819	–	.1699	.2732	.5082
n target docs	$n = 3$	.3907	.2227	.4654	$n = 10$	.3036$^{\blacktriangle}$	.3854	.5705
n wikipedia docs	$n = 2$	.3964	.2136	.4425	$n = 7$	.4266$^{\blacktriangle}$	.4537$^{\blacktriangle}$	.6988$^{\triangle}$
n source docs	$n = 7$	.4949$^{\triangle}$	.2818	.5435	$n = 5$	.4988$^{\blacktriangle}$	.4829$^{\blacktriangle}$	.6864$^{\triangle}$

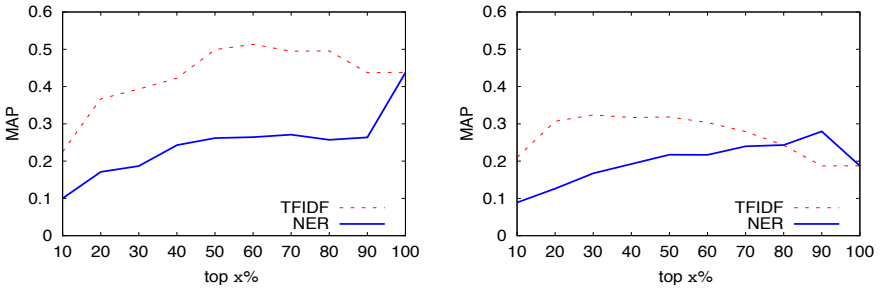
Table 4. Results of the term selection experiments; significance tested against the baseline

	Same event				Related event			
TS Model	*detail*	*MAP*	*P5*	*MRR*	*detail*	*MAP*	*P5*	*MRR*
baseline	–	.3623	.2227	.4820	–	.1699	.2732	.5083
content	–	.3582	.1955	.4800	–	.1583	.2634	.4838
metadata	–	.1636$^{\blacktriangledown}$	.0636$^{\blacktriangledown}$	.1863$^{\blacktriangledown}$	–	.1768	.2000	.2887$^{\blacktriangledown}$
title	–	.4157	.2227	.4597	–	.2264	.2829	.4300
lead	–	.4428	.2318	.5386	–	.2681	.3366	.5294
$x\%$ *terms*	$x = 60\%$	.5133$^{\triangle}$	.2682	.6390	$x = 30\%$	.3229	.3268	.4799
$y\%$ *ne*	$y = 100\%$	.4374	.2091	.5592	$y = 90\%$	.2796	.2829	.4724
combined	x=60%, y=100%	.4660	.2409	.5849	x=30%, y=90%	.3387	.3317	.4459

expansion works as it adds text from multiple news articles about the parliamentary discussion on transparency to the target item, compensating for different views. Similarly, for *related event linking*, document expansion increases the number of viewpoints of a seminal event covered in a source item to improve linking performance.

5.2 Term Selection

On the *source* item side we experiment with different term selection techniques. In this section, we link to the original unexpanded target items. Table 4 shows that using only terms from a specific field, e.g., lead or title, improves over using the whole document in terms of absolute scores for both *same event linking* and *related event linking*, but not significantly so. We also select terms and named entities from the content of the *source* item based on their TFIDF score. Figure 3a shows the MAP score for *same event linking* while using only the top $x\%$ of the terms (dotted line) or named entities (solid line). We observe that removing any named entities decreases performance. For selecting terms there is an optimum when only 60% of the terms (ranked by TFIDF) are selected. Table 4 shows that *same event linking* with the optimum of 60% of the terms selected from the source item, a significant improvement over the baseline is achieved. When linking to related events, selecting terms from the source item does not lead to significant improvements over the baseline; this is not surprising as *related event linking* is more recall oriented and benefits from having a source item description that covers



(a) Same event **(b)** Related event

Fig. 3. Term selection with the top $x\%$ (ranked by TFIDF) of the original terms and named entities from the original source item retained

all aspects of a seminal event. When less terms from the source item are selected, less aspects of an event are covered making linking to related events more difficult.

We take a closer look at the source item that benefits most from term selection on the *same event linking* task. The title of the source item is "Ukrainian president dissolves parliament." One of the target items is described by: "President Yushchenko of Ukraine dissolves parliament and issues new elections." The source item content consists of 342 words and mentions various aspects of the event, e.g., comments of the opposition leader and protests leading to dissolution of parliament. Each aspect potentially matches with the description of a target item. As the political situation in the Ukraine was unstable for a number of years, many target items cover aspects of this topic. By only selecting a small number of terms specific to the seminal event, term selection prevents a drift in topic towards other aspects of the source item description.

5.3 Further Improving Linking Performance

In order to see how far we can push linking performance we conduct two additional experiments. In the first we combine the best models, i.e., the best term selection is used to find targets and the target items have been expanded with the optimal number of documents. The combination achieves a MAP of .4801 on the *same event linking* task, which does not improve over using document expansion (.4949) or term selection (.5133) by itself. We find similar results for *related event linking*. We find that for items where document expansion helps, term selection has relatively poor performance, and vice versa. This fits the intuition that term selection and expansion have opposite effects: one makes an item's event description more specific, while the other broadens the description. Depending on the source item only one of the effects may be desired.

Our second experiment is with a date filter that restricts target items to a period of 14 days around the date of the source item. On the *same event linking* task this results in a baseline MAP score of .5689 and scores of .7263 and .7397 MAP for the best document expansion and term selection models, respectively. Scores for all models go up, including the baseline, but the same significant differences in performance remain between the baseline and the best models. On the *related event linking* task using a date filter decreases performance, from .1699 to .0883 MAP for the baseline and to .1487

MAP and .2077 MAP for the optimal document expansion and term selection settings, respectively. That filtering on dates improves performance on *same event linking* is unsurprising: this is a high precision-oriented task and a news broadcast and article about the same event are published around the same day. This effect, however, is specific to linking news based on same events. Related events, which may be distributed over a long period, do not demonstrate the same effect.

6 Conclusions

With archives opening up their content online and enabling interconnectivity, the challenge of linking between items in different archives arises. We consider the task of linking archives based on events. We investigate two variants of this task: *same event linking* and *related event linking*. We use a retrieval approach to link items from a news paper archive with very rich text descriptions to videos in a multimedia archive with relatively sparse annotations. This mismatch between the representations in both archives gives rise to our two research questions: (i) does expanding sparse item representations with text from other sources improve linking performance; and (ii) what effect does term selection on a textually rich source item have on linking performance?

In answer to (i), we find that expanding *target* items with documents from other sources improves performance for both *same event linking* and *related event linking*. Using expansion documents from the source archive, however, is most effective as the content has the same focus as the target archive. In answer to (ii), we find that reducing the number of terms in the *source* item representation is most effective for *same event linking*. The reduced items are more robust to topic drift and form a better match for the short event descriptions in the target archive. *Related event linking* also improves but not as much as with target item expansion. Related events benefit more from rich descriptions (as obtained through expansion) that cover all aspects of an event.

Turning to directions for future work, our combination of document expansion and term selection techniques would benefit from further investigation. We combined the best document expansion model with the best term selection settings, but this naive approach did not outperform the individual methods. Another direction is investigating other settings where rich and sparse archives need to be linked; our document expansion and term selection techniques are general enough to be applied in non-news settings, e.g., linking art encyclopedias (with rich text) to museum collections (with sparse text).

Acknowledgements. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

References

[1] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y., et al.: Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218 (1998)

[2] Bron, M., van Gorp, J., Nack, F., de Rijke, M.: Exploratory search in an audio-visual archive: Evaluating a professional search tool for non-professional users. In: EuroHCIR 2011: 1st European Workshop on Human-Computer Interaction and Information Retrieval (July 2011)

[3] Carrick, C., Watters, C.: Automatic association of news items. Information Processing & Management 33(5), 615–632 (1997)

[4] Cohn, D., Hofmann, T.: The missing link-a probabilistic model of document content and hypertext connectivity. In: NIPS 2001, pp. 430–436 (2001)

[5] Diaz, F., Metzler, D.: Improving the estimation of relevance models using large external corpora. In: SIGIF 2006, pp. 154–161. ACM, New York (2006)

[6] Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: ACL 2005, pp. 363–370. ACL (2005)

[7] Franz, M., Ward, T., McCarley, J., Zhu, W.: Unsupervised and supervised clustering for topic tracking. In: SIGIR 2001, pp. 310–317. ACM, New York (2001)

[8] Harman, D.K.: The TREC test collections. In: Voorhees, E.M., Harman, D.K. (eds.) TREC: Experiment and Evaluation in Information Retrieval. MIT, Cambridge (2005)

[9] Henzinger, M., Chang, B.-W., Milch, B., Brin, S.: Query-free news search. In: World Wide Web, vol. 8, pp. 101–126 (2005)

[10] Huurnink, B., Hollink, L., van den Heuvel, W., de Rijke, M.: Search behavior of media professionals at an audiovisual archive: A transaction log analysis. J. American Soc. Information Science and Technology 61(6), 1180–1197 (2010)

[11] Kern, R., Granitzer, M.: German encyclopedia alignment based on information retrieval techniques. In: Lalmas, M., Jose, J., Rauber, A., Sebastiani, F., Frommholz, I. (eds.) ECDL 2010. LNCS, vol. 6273, pp. 315–326. Springer, Heidelberg (2010)

[12] Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: SIGIR 2004, pp. 297–304. ACM, New York (2004)

[13] Li, Z., Wang, B., Li, M., Ma, W.: A probabilistic model for retrospective news event detection. In: SIGIR 2005, pp. 106–113. ACM, New York (2005)

[14] Ma, Q., Nadamoto, A., Tanaka, K.: Complementary information retrieval for cross-media news content. Information Systems 31(7), 659–678 (2006)

[15] Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Learning semantic query suggestions. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 424–440. Springer, Heidelberg (2009)

[16] Mihalcea, R., Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In: CIKM 2007, vol. 7, pp. 233–242 (2007)

[17] Radev, D., Otterbacher, J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence: summarizing online news topics. Comm. of the ACM 48(10), 95–98 (2005)

[18] Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Comm. of the ACM 18(11), 613–620 (1975)

[19] Tao, T., Wang, X., Mei, Q., Zhai, C.: Language model information retrieval with document expansion. In: HLT-NAACL 2006, pp. 407–414 (2006)

[20] Tsagkias, M., de Rijke, M., Weerkamp, W.: Linking online news and social media. In: WSDM 2011, pp. 565–574. ACM, New York (2011)

[21] Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: SIGIR 2002, pp. 81–88. ACM, New York (2002)

Transformation of a Keyword Indexed Collection into a Semantic Repository: Applicability to the Urban Domain

Javier Lacasta[1], Javier Nogueras-Iso[1], Jacques Teller[2], and Gilles Falquet[3]

[1]Computer Science and Systems Engineering Dept. Universidad de Zaragoza, Spain
[2]LEMA, Université de Liège, Belgium
[3]Centre universitaire d'informatique, Université de Genève, Switzerland

Abstract. In the information retrieval context, resource collections are frequently classified using thesauri. However, the limited semantics provided by thesauri restricts the collection search and browsing capabilities. This work focuses on improving these capabilities by transforming a set of resources indexed according to a thesaurus into a semantically tagged collection. The core mechanism for building this collection is based on the conversion of the domain specific thesaurus (indexing the collection of resources) into a domain ontology connected to an upper level ontology. The feasibility of this work has been tested in the urban domain by transforming the resources accessible through the European Urban Knowledge Network into a Linked Data repository.

1 Introduction

In the information retrieval context, the resources of a collection are frequently classified and searched using the concepts of thesauri. However, the limited semantics they provide reduce its usability for search and browsing in a collection. Thesauri can be used to expand queries to a collection by including narrower concepts of the selected one (they are about the same theme), but only in a rough way since the lack of semantics in their relations increase the heterogeneity of criteria and interpretations.

With the objective of improving the search and browsing capabilities of a collection, this paper proposes a process to convert a thesaurus indexed collection into a semantically tagged collection stored in a semantic repository. The process to perform this transformation is based on the transformation of the thesaurus used to classify the collection into an ontology. Formal ontologies provide a more detailed structure with improved reasoning capabilities. They describe in detail the meaning of each of the included concepts and the specific types of relations held between them. However, it is a waste of effort to create new models from scratch if a thesaurus already exists in the desired application domain. It is much more suitable to convert them into an ontology by adding the semantics they lack.

The mechanism proposed for the transformation of a thesaurus into an ontology is based on linking it to DOLCE ontology [10] and using this linkage to refine

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 372–383, 2011.

and extend the original thesaurus model. The method is applied in the urban domain to transform the EUKN resources[1] into a semantically tagged collection. The EUKN thesaurus has been formalized following the described process and the collection has been republished using a Linked Data web service. Linked Data models provide RDF resources and relations between them as valid HTTP URIs, facilitating in that way the access and browsing through the information. This transformation provides, in addition to the thematic entry point to the collection, a more abstract point of view focused on DOLCE categories (activities, events, rational-agents ...) and a set of extended relations between the resources.

The rest of the paper is structured as follows. Section 2 reviews the state of the art. Section 3 introduces the proposed method. Section 4 shows the applicability of the method to the urban domain. Finally, this paper ends with some conclusions and outlook for future work.

2 State of the Art in the Generation of Semantic Models

This section reviews the main works dealing with the addition of semantics to collections of resources. It analyzes the works focused on converting resource descriptions into semantic networks, and those related to the formalization of the terminological models used in their classification.

With respect to the conversion of sets of resources into semantic networks, Hearst et al. [5] review the existing approaches. They describe alternatives for searching and browsing collections focusing on the use of faceted search/browsing components based in controlled knowledge models to help to guide the user in the location of the desired resources. In this same field, Hyvöenen [6] describes a content creation process that includes the manual transformation of the thesaurus used for classification of a collection (*Museoalan asiasanasto*) into an ontology and the generation of the collection records in RDF.

Without the final objective of transforming a set of resources into a semantic model (but that can also be applied for this task), there are works that analyze other alternatives for the formalization of thesaurus. In this area of work, Tudhope et al. [13] analyze the specialization of the associative thesaurus relations into richer subtypes to find new application possibilities for retrieval. Similarly, Golbeck et al. [3] describe the process used to transform the National Cancer Institute (NCI) thesaurus into an ontology. They show the rules applied to transform each concept, property and relation of the thesaurus into formal equivalents. Other works go a bit further by proposing semiautomatic processes to perform these formalization tasks. Kawtrakul et al. [8] and Soergel et al. [12] present a (semi-)automatic process to refine the relations of the AGROVOC thesaurus based on the analysis of the AGROVOC categories that classify the thesaurus concepts. They are used to establish abstract relations that are applied as general transformation rules for their member relations. In a similar way, Khosravi and Vazifedoost [9] propose a re-engineering process of the ASFA

[1] http://www.eukn.org/E_library

Persian thesaurus using automatic ontology learning methods. It is based on the definition (by experts in the field) of general and specific rules used to transform the thesaurus relations.

The need of formalization is not restricted to thesauri. Collections using other semi-structured or unstructured knowledge models for classification also benefit from the formalization of the used models. For example, Van Damme et al. [2] show how folksonomies and other unstructured vocabularies can be used to construct ontologies. They describe a multiple approach for deriving ontologies from folksonomies, such as the statistical analysis of the folksonomies, the use of online lexical and semantic web resources, ontology matching (and mapping) approaches, and the computer guided human review. Finally, focusing on controlled lists of terms, Aleksovski et al. [1] propose a method to match two unstructured lists of terms through a background ontology using different disambiguation and heuristic techniques.

3 Providing Semantics to a Keyword Indexed Collection

The proposed method for adding semantics to a collection of resources is described in figure 1.

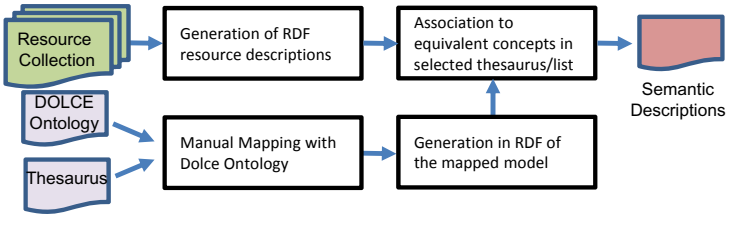


Fig. 1. Generation process of a semantic description of a resource collection

The first step is the generation in RDF of the resource descriptions (or to process them into a structured model if they already exist). This step is dependent on the original source (e.g., HTML, XML, word documents...) and the techniques and processes required to do that can be completely different. The objective is to obtain for each resource a Dublin Core [7] metadata record expressed in RDF containing at least: the resource name, a description, a set of keywords that classify them, and a reference to the original resource. The Dublin Core metadata model was selected for resource description because its extensive use in the digital library field facilitates the access to resources from a wide audience. Next, those fields whose values can be linked to a knowledge model (e.g., keywords, authors, dates, locations) are processed to convert them into independent RDF items linked to Dublin Core RDF resources. Due to lexical heterogeneity of labels (e.g., plurals, misspellings errors...), in order to avoid the creation of duplicated resources, it is needed to harmonize the labels. Figure

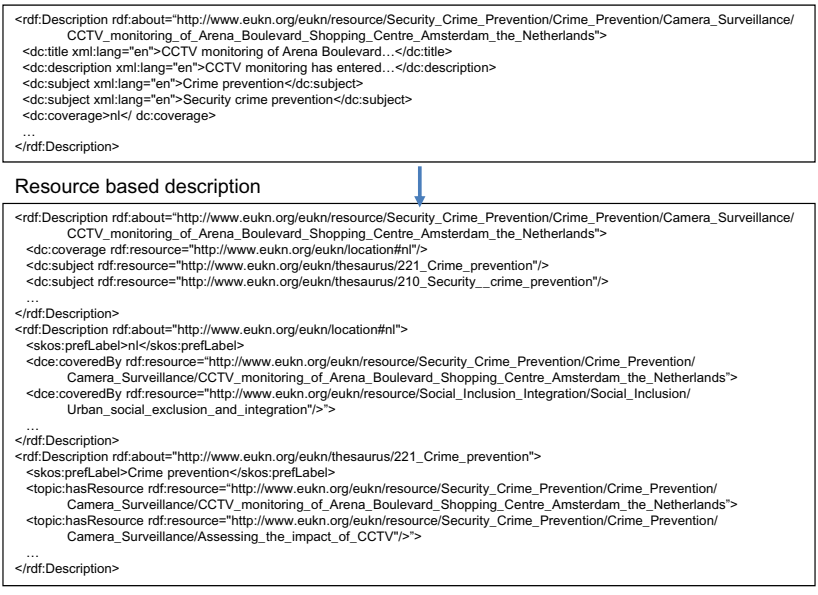


Fig. 2. Example of transformation of properties into resources

2 exemplifies how Dublin Core metadata containing controlled values can be transformed into independent resources. Additionally, the requirements of the library used later in the construction of a Linked Data based system force all the relations between resources to be bidirectional (i.e., they have an inverse).

Once the descriptions of the resources have been processed, the next step is to replace their (string) subjects with equivalent concepts from well-known knowledge models (e.g., a thesaurus) that provide a better organization and description of concepts. The required matching process is similar to the one performed to identify equivalent properties in different Dublin Core descriptions. The main difference comes from the existence in the knowledge models of different alternative names (synonyms), descriptions, and similarity relations that can be used to help in the matching process. This improves the collection organization; however, the semantics provided by a simple knowledge models is limited. If no description or scope note is provided, a concept meaning can be ambiguous. Additionally, the hierarchical relations contained in thesauri only indicate a general meaning containment and the associative ones only provide an abstract level of relatedness. The formalization of these models helps to identify each concept meaning and the relations between them.

The approach adopted to formalize a simple knowledge model has been to interrelate it with other already existent formal models. As base for the formalization process it has been decided to use a top level ontology such as DOCLE, a model focused on describing data types and general relations independent of the context [4]. Mika et al [11] describe some reasons why knowledge models should be linked to a top level ontology. On the one hand, it provides additional seman-

tics about the concepts and relations to determine if the model is coherent. On the other hand, it aggregates the concepts and relations into abstract categories that can be used to automate the establishment of domain oriented relations. The objective here is to link the thesaurus with the top level ontology through *subclass-of* relations. Using DOLCE as top level ontology, it is possible to perform a manual matching process between the desired thesaurus concepts and the DOLCE classes contained in the *perdurant*, *endurant*, and *quality* branches. *Perdurants* comprise events, processes, phenomena, activities and states; *endurants* describe entities that maintain their identity along the time, although their properties (e.g., color, size...) may change; finally, *qualities* provide entities that can be perceived or measured (e.g., shape, color).

Once the matching has been performed, the relations between the concepts in the original model (e.g., hierarchical and associative relations) can be automatically refined through inference of the corresponding relations in the mapped ontology. For example, in DOLCE ontology it is described that two *physical-objects* may hold a *part-of* relation. Therefore, *ICT infrastructure* is a narrower of *Technical infrastructure* and those concepts can be classified as DOLCE *physical-objects*, it can be automatically deduced that they hold a *part-of* relation between them.

The resulting model containing the collection resources linked to a formalized thesaurus can be directly stored in a semantic repository such as SESAME or JENA to allow the execution of semantically rich queries and the creation of faceted guided search systems. For example, the DOLCE enriched ontology can be used to search activities, rational-agents, regulations and so on. The publication of the collection as Linked Data is immediate once it is stored in a semantic repository. Tools as PUBBY allow the creation of Linked Data services from a semantic model by transforming the URIs of the resources into valid URLs.

4 Applicability of the Method to the EUKN Collection

The described process has been applied to the European Urban Knowledge Network (EUKN) collection with the objective of adding semantics to it and providing alternatives for search and browsing. The following subsections describe: the features of the EUKN collection and its thesaurus; the results of applying the transformation method to link the EUKN thesaurus to an upper level ontology; and the publishing process of the EUKN collection as Linked Data together with a discussion of the advantages of the new approach for searching and browsing the collection.

4.1 The EUKN Collection and Its Thesaurus

The European Urban Knowledge Network (EUKN) was born in 2004 as a pilot project of different European states to enhance the exchange of knowledge and expertise on urban development. Nowadays, it is an intergovernmental knowledge network that acts as hub for existing networks of urban practitioners, researchers

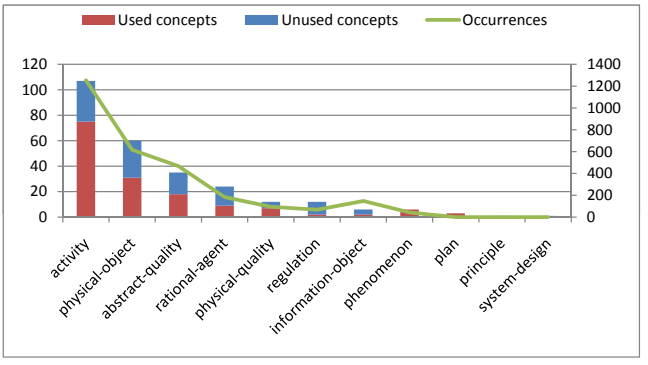


Fig. 3. Analysis of the use of EUKN concepts matched with DOLCE categories for the classification of resources in the collection

and policy-makers at all governmental levels. It provides a high-quality knowledge database in the urban field with more than 2,000 resources classified according to a thesaurus that contains 267 concepts to describe the management and control of physical, economic and social forces on urban areas. However, there is a disconnection between the thesaurus creators and their users because only 146 (54%) of those concepts are used in the collection. It may be caused by structural deficits that difficult the location, identification, and usability of the thesaurus concepts. For example, it contains some replicated concepts, the concepts lack a detailed definition and/or scope notes, and the criteria for its hierarchical structure is sometimes unclear (e.g., *Mediation* as narrower of *Community development*).

4.2 Transformation of the EUKN Thesaurus into an Ontology

The transformation of the EUKN thesaurus into an ontology has been performed linking it to DOLCE top level ontology as described in section 3. The resulting mapping uses 18 DOLCE classes (no more classes have been needed). However, in order to minimize the heterogeneity, the classes being very similar or with few associated concepts (e.g., *planning activities* or *geographical-objects*) have been integrated as part of their corresponding super-classes.

Figure 3 analyzes the results of the thematic association of DOLCE categories with the EUKN concepts. It shows the number of thesaurus concepts associated to each DOLCE category and how the concepts in these clusters are used in the EUKN collection. We have obtained the following classification: 107 *activities*, 61 *physical-objects*, 45 *abstract-quality*, 24 *rational-agents*, 18 *regulations*, 7 *information-objects*, and 6 *phenomenon*. The red section of each bar (Used concepts) identifies the number of concepts classified in that category and used in the EUKN collection. The blue one (Unused concepts) identifies those that have not been used in the classification. Finally, the graph line (Occurrences) shows the number of EUKN resources classified using concepts of each

category (right axis). It can be observed that the thesaurus is mainly focused about the activities, elements and characteristics related to urbanism with other lesser themes such as the persons, groups and regulations involved in it. Only two minor quality issues have been identified. Firstly, the number of resources focusing on *abstract-qualities* and in *information-objects* is too high for the weight that it has been given in the thesaurus. This means a fewer degree of discrimination in searches about these fields. Secondly, some of the identified categories in DOLCE have very few associated concepts and are not even used in the EUKN collection. The first issue is a conceptual problem that would require a restructuring in the next thesaurus version. The second one can be directly fixed by removing the unused leafs to adjust the thesaurus model to the use it has in the collection.

Once the EUKN thesaurus concepts have been linked to a DOLCE class, the next step has been to redefine their broader/narrower relations. This has been done by replacing them with an existing relation between their associated DOLCE classes. However, since DOLCE may provide several possible relations between two classes (e.g., the relation between a geographical-object and a physical-object can be of type *part* or of type *subclass*), a set of rules have been defined to select the most suitable one for each case and apply it to all equivalent cases. For example, if A is the *narrower* of B in the EUKN with A being a *physical-object* and B a *abstract quality*, the *narrower* should be transformed into a *has-quality* relation. Table 1 shows the rules to infer the DOLCE relations that replace the original broader/narrower relations of the thesaurus on the basis of the associated DOLCE classes.

An example of relations refinement is shown in figure 4. It shows how the 10 narrower relations of the *Environmental sustainability* branch have been processed. In the classification process, the *Environmental sustainability* concept and three of its children have been tagged as a DOLCE *activity*; the other 7 have been classified as DOLCE *physical-quality*. Following the rules indicated in table 1, the narrower relation between two *activities* must be replaced with the DOLCE *result-of* relation. In the case of narrower relations between an *activity* and a *physical-quality*, the *has-quality* relation has been used. With this

Table 1. Inferred relation

Pairs of DOLCE classes associated with EUKN concepts	**Relation**
(activity → physical/abstract-quality) (geographical/physical/information-object → abstract-quality) (rational-agent → abstract-quality) (regulation → abstract-quality) (plan → abstract-quality) (physical-quality → abstract-quality) (physical-quality → physical-quality)	has-quality
(activity → rational-agent) (activity → information/physical-object) (activity → regulation) (activity → principle) (phenomenon → geographic-object)	participant
(abstract-quality → abstract-quality) (activity → plan) (phenomenon → activity) (geographic-object → geographic-object) (regulation → plan)	part
(plan → activity) (rational-agent → information-object) (rational-agent → physical-object) (rational-agent → plan) (norm → system-design)	generic-dependent
(geographical-object → physical-object) (rational-agent → rational-agent) (regulation → regulation) (information-object → information-object)	subclass-of
(physical-object → activity) (physical-object → plan)	instrument-of
(activity → activity)	result-of

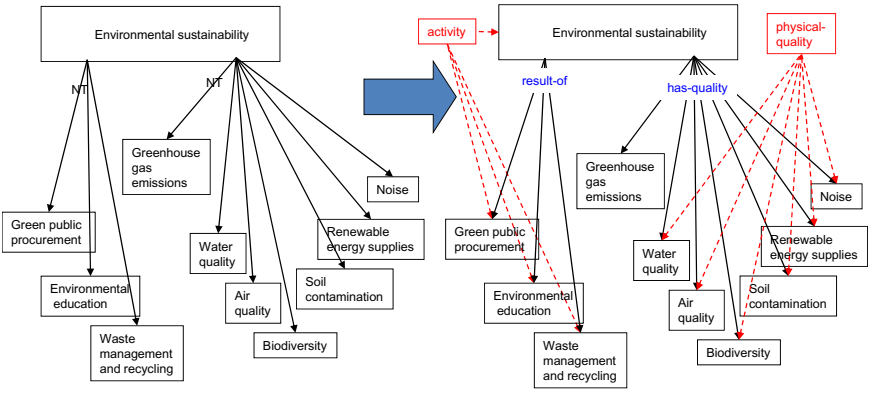


Fig. 4. Transformation of the *Environmental sustainability* concept and its narrower concepts

improved model, an additional level of knowledge is obtained. Now, it is possible to infer that *Environmental sustainability* is the *result-of Waste management and recycling*, *Environmental education*, and *Green public procurement*. Additionally, it can be measured (*has-quality*) through the *Water quality*, the *Air quality* and so on.

4.3 Publication of the EUKN Collection

The storage of the EUKN collection in a semantic repository and its publication as linked data requires a complete transformation format and the connection with the conceptual model created previously. The first step has been to process the HTML resources and convert them into a Dublin Core RDF model. In this model, information such as the title or description has been stored as property values. However, those fields whose values are shared by several records (e.g., authors, organizations, document types, or coverage) are stored as URIs referencing to independent resources containing the shared information. The thematic information is not replaced with references to new resources, but to the corresponding ones in the knowledge model developed previously. If other fields with replicated values (e.g., the locations, authors or organizations) were also modeled with an ontology, the same association should be done for them (to improve the model quality). Figure 5 shows how a EUKN resource is represented and how it is bidirectionally associated to the *Biodiversity* concept of the transformed EUKN ontology through the Dublin Core *subject* property.

The integrated model has been then stored in a JENA semantic repository and accessed through an SPARQL end point provided by the JOSEKI library. The SPARQL endpoint provides the dual functionality to facilitate an open query system for advanced users and as base for the construction of simpler and specific query and browsing components. A thematic graphical query component that transforms the user interaction in corresponding SPARQL queries

```
<rdf:Description
rdf:about="http://www.eukn.org/eukn/resource/Urban_Environment/Environmental_Sustainability/
        Biodiversity/Urbanisation_can_be_an_opportunity_or_a_threat_for_biodiversity">
    <dc:title xml:lang="en">Urbanisation can be an opportunity or a threat ...</dc:title>
    <dc:subject rdf:resource="http://www.eukn.org/eukn/thesaurus/11_Biodiversity"/>
    <dc:coverage rdf:resource="http://www.eukn.org/eukn/location#eu"/>
    <dc:description xml:lang="en">The report '10 messages for 2010 - Urban Ecosystems',
       published by the European Environment Agency (EEA), provides an overview of the
       relation between urban ecosystems and biodiversity </dc:description> ...
</rdf:Description>

<rdf:Description rdf:about="http://www.eukn.org/eukn/thesaurus/11_Biodiversity">
    <rdfs:subClassOf rdf:resource=
        "http://www.eukn.org/eukn/thesaurus/dolceEq#physical-quality"/>
    <dolce:inherent-in rdf:resource=
        "http://www.eukn.org/eukn/thesaurus/9_Environmental_sustainability"/>
    <topic:hasResource rdf:resource="http://www.eukn.org/eukn/resource/Urban_Environment/
        Environmental_Sustainability/Biodiversity/
        Urbanisation_can_be_an_opportunity_or_a_threat_for_biodiversity"/>
    <skos:prefLabel xml:lang="en">Biodiversity</skos:prefLabel> ...
</rdf:Description>
```

Fig. 5. Example of RDF generated for a resource

to the endpoint has been developed. With respect to the browsing through the collection, the Linked Data service PUBBY has been used. PUBBY performs the transformation between the URIs used to link the resources in the collection and valid URLs that provide the desired resource concepts in the web. This provides a simple way to browse an RDF collection but at the cost of using for browsing a set of URIs that are different from the contained in the repository and accessed through the SPARQL endpoint. This issue will have to be dealt to provide homogeneous URIs from both services.

Figure 6 shows three examples of SPARQL queries. The first one demonstrates that thanks to the development of this repository, it is now possible to search resources based on their metadata descriptions (the original collection of HTML records describing the resources did not allow field based search functionality). The second one shows how complex queries can be easily expressed in SPARQL. In particular, it shows how to retrieve all the resources annotated with concepts being part of an urban technical infrastructure. Using a traditional digital library system, we should have first expanded our query to include all the narrower terms of urban technical infrastructure, and later search the metadata database. Finally, the third example shows the potential of inference reasoners. It returns all resources annotated with EUKN concepts directly classified as DOLCE activities (concepts described as subclass of activity), or subclasses of these EUKN concepts.

With respect to the browsing, the thematic search components provide a first entry point to the desired resources in the collection. The original EUKN thesaurus is provided as a tree in which the selection of a term returns the resources classified according to that term, but also (if requested) it can return those classified according to all their narrowers. Additionally, it shows the enriched EUKN thesaurus model with the DOLCE categories used to organize the concepts and allows browsing through the updated relations. Additionally, the linked visu-

```
Select ?resUri where { ?resource dc:source ?resUri.
       ?resource dc:title ?title. FILTER regex(?title, "town", "i") }
Select ?title where { ?concept topic:hasResource ?resource. ?resource dc:title ?title.
       ?concept dolce:part-of <http://www.eukn.org/eukn/thesaurus/90_Technical_infrastructure>}
Select ?title where { ?resource dc:title ?title.
       <http://www.eukn.org/eukn/thesaurus/93_Electricity> topic:hasResource ?resource.
       ?concept rdfs:subClassOf dolce:activity. ?concept topic:hasResource ?resource}
```

Fig. 6. SPARQL query examples

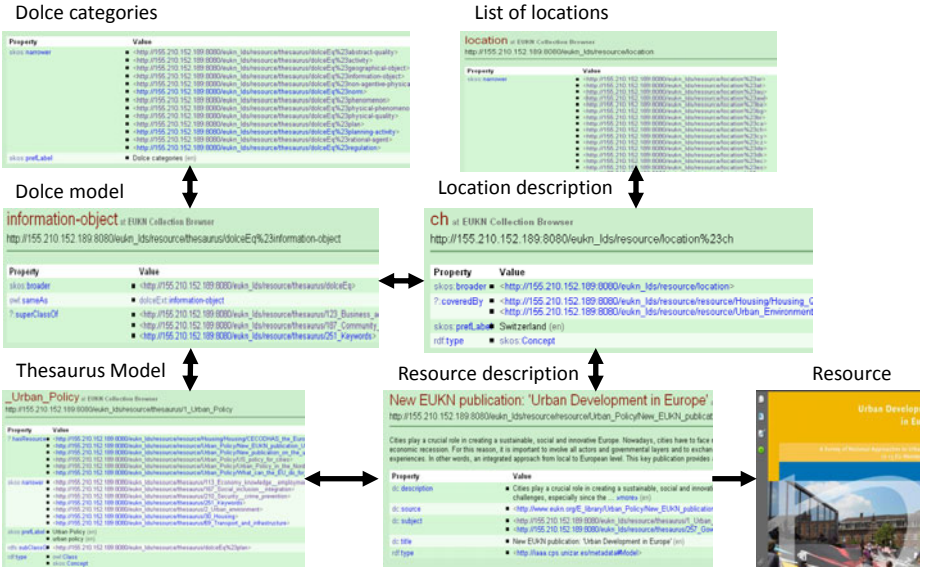


Fig. 7. Overview of the browsing system

alization facilitates the access to the information providing a rich structure of related resources. Figure 7 shows snapshots of the browsing system[2] and the links that hold between them. It can be observed how in addition to the EUKN thesaurus based navigation, browsing based on DOLCE categories and in other controlled fields is also possible.

This system provides an improvement with respect to the original one in the sense that it facilitates guided multi criteria search and browsing through the collection through a conceptual view of the collection instead of (or combined with) a thematic one. Experienced users maintain the classical thematic access to the collection but with an improved model and more precise relations between the concepts. Additionally, the collection can also be accessed through DOLCE categories providing a more generic conceptual access.

[2] http://mularroya04.cps.unizar.es:8080/eukn_sparql/

5 Conclusions

This paper has described a process for the transformation of a collection of resources indexed with a thematic thesaurus (and described mostly as free text) into a semantically tagged collection that can be accessed and browsed as Linked Data.

The process described is focused on two tasks: enriching the thesaurus used for the collection classification using a top level ontology such as DOLCE, and transforming the textual resource collection into a set of interrelated RDF resources. The enrichment is based on a semiautomatic matching process between the thesaurus and DOLCE where the relations are automatically inferred. The transformation of the collection resources into RDF is left open due to the dependence of the specific collection to process.

The process has been applied to the EUKN collection. The resulting classification ontology and enriched collection has been stored in a JENA semantic repository and accessed through a facet-based search system allowing the browsing through the collection using a PUBBY Linked Data service. The search system provides access to the collection through the original and the enriched thesaurus. It maintains the original access for experience users, but it also offers a conceptual entry-point and collection browsing for inexperience ones.

Future work will firstly focus on validating the generated ontology to measure its quality and correct possible imprecisions in the established relations. When the model is verified, the system is expected to be published and improved through the obtained feedback. Additionally, we want to explore how the mapping with DOLCE can help to measure the thesaurus quality, and improve it, if needed. For example, very heterogeneous categories with few members may indicate a poor concept selection or a dispersion in the thesaurus classification objectives. Moreover, relations between concepts whose DOLCE equivalents do not hold a suitable relation may indicate an organization error in the thesaurus concept structure.

With respect to the browsing system, future work will focus on the improvement of the faceted system by providing additional access characteristics such as locations, authors, organizations or dates. We will need to model how each of these elements are organized and create a suitable ontology for their domain values. This will expand the resulting linked model enhancing their capabilities. In this context, the user interface must be improved to show human friendly labels (currently it shows URIs) sorted in an appropriate way (e.g., alphabetically).

Acknowledgements. This work has been partially supported by Spanish Government by the projects TIN2009-10971 and "España Virtual" ref CENIT 2008-1030 (through contracts with the National Center of Geographic Information and GeoSpatiumLab), Geo-SpatiumLab S.L. and Zeta Amaltea S.L.

References

1. Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F.: Matching unstructured vocabularies using a background ontology. In: Staab, S., Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 182–197. Springer, Heidelberg (2006)

2. Damme, C.V., Hepp, M., Siorpaes, K.: Folksontology: An integrated approach for turning folksonomies into ontologies. In: Bridging the Gap between Semantic Web and Web 2.0 Workshop, ESWC 2007, Innsbruck, Austria, pp. 1–14 (2007)
3. Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Parsia, B., Oberthaler, J.: The national cancer institute's thesaurus and ontology. Journal of Web Semantics 1(1), 1–5 (2003)
4. Guarino, N.: Formal Ontologies and Information Systems. In: Amsterdam, I.P. (ed.) Proceedings of FOIS 1998, Trento, Italy, pp. 3–15 (June 1998)
5. Hearst, M.: Design Recommendations for Hierarchical Faceted Search Interfaces. In: ACM SIGIR Workshop on Faceted Search (August 2006)
6. Hyvönen, E., Salminen, M., Junnila, M., Kettula, S.: A content creation process for the semantic web. In: Proceeding of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments (May 2004)
7. International Organization for Standardization: Information and documentation - The Dublin Core metadata element set. ISO 15836:2003, International Organization for Standardization (ISO) (November 2003)
8. Kawtrakul, A., Imsombut, A., Thunkijjanukit, A., Soergel, D., Liang, A., Sini, M., Johannsen, G., Keizer, J.: Automatic Term Relationship Cleaning and Refinement for AGROVOC. In: Workshop on The Sixth Agricultural Ontology Service, Vila Real, Portugal (July 2005)
9. Khosravi, F., Vazifedoost, A.: Creating a persian ontology through thesaurus reengineering for organizing the digital library of the national library of iran. In: International Conference on Libraries, Information and Society, ICoLIS 2007, Malaysia, pp. 19–36 (June 2007)
10. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L.: Wonderweb deliverable d17: The wonderweb library of foundational ontologies. Tech. rep., ISTC-CNR (2003)
11. Mika, P., Oberle, D., Sabou, M., Gangemi, A.: Foundations for service ontologies: Aligning owl-s to dolce alignment to foundational ontologies. In: Proceedings of the Thirteenth International World Wide Web Conference, New York, USA (May 2004)
12. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering Thesauri for New Applications: the AGROVOC Example. Journal of Digital Information 4(4), 1–19 (2004)
13. Tudhope, D., Alani, H., Jones, C.: Augmenting Thesaurus Relationships: Possibilities for Retrieval. Journal of Digital Information 1(8), 22 pages (2001)

Improving Europeana Search Experience Using Query Logs

Diego Ceccarelli[1,2], Sergiu Gordea[3], Claudio Lucchese[1], Franco Maria Nardini[1], and Gabriele Tolomei[1,4]

[1] ISTI–CNR, Pisa, Italy
{name.surname}@isti.cnr.it
[2] Dipartimento di Informatica, Università di Pisa, Italy
[3] AIT Austrian Institute of Technology GmbH, Wien, Austria
{name.surname}@ait.ac.at
[4] Università Ca' Foscari, Venezia, Italy

Abstract. Europeana is a long-term project funded by the European Commission with the goal of making Europe's cultural and scientific heritage accessible to the public. Since 2008, about 1500 institutions have contributed to Europeana, enabling people to explore the digital resources of Europe's museums, libraries and archives. The huge amount of collected multi-lingual multi-media data is made available today through the Europeana portal, a search engine allowing users to explore such content through textual queries. One of the most important techniques for enhancing users search experience in large information spaces, is the exploitation of the knowledge contained in query logs. In this paper we present a characterization of the Europeana query log, showing statistics on common behavioral patterns of the Europeana users. Our analysis highlights some significative differences between the Europeana query log and the historical data collected by general purpose Web Search Engine logs. In particular, we find out that both query and search session distributions show different behaviors. Finally, we use this information for designing a query recommendation technique having the goal of enhancing the functionality of the Europeana portal.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 384–395, 2011.

1 Introduction

The strong inclination for culture and beauty in Europe created invaluable artifacts starting from antiquity up to nowadays. That cultural strength is recognized by all people in the world and makes Europe the destination for a half of the international tourists[1]. More than 220 million people visit the European countries yearly for spending their holidays.

The European Commission is aware about the value of this cultural heritage and decided to make it more accessible to the public by supporting digitization of the cultural heritage and by financing the Europeana group projects. The first prototype of the Europeana Portal[2] was launched in autumn 2008 and contains by now about 15 million items.

Due to increasing amount of information published within the portal, the access to the description of a specific masterpiece becomes each day a more time consuming task, when the user is not able to create a very restrictive query. For example, if we search today in Europeana for general terms like *renaissance* or *art nouveau* we will find more than 10,000 results. If we search for the term *Gioconda* we find a couple of hundred of items, and if we search for *Mona Lisa, Da Vinci* we get 20 images of the well known painting. These examples show how important is to use good queries when looking for very particular information on the web by using a search engine like Europeana. This is a challenging task, given the fact that the document base is cross-domain, multi-lingual and multi-cultural.

Search query recommendation techniques [3,11] are commonly used in web search engines to help users to refine their queries. These technologies analyze the user behavior by mining the system logs in order to find the correlation between what the user's information need (visited pages), what the user is searching for (query terms) and the content and structure of the information pool (search index).

In this paper we present the work carried out by now in the ASSETS[3] project with the goal of implementing a query recommendation module for Europeana port. We focus our attention on the analysis of the user behavior and particularities of the information pool.

The rest of the paper is organized as follows: Section 2 introduces related work, while Section 3 discusses the main results coming from the analysis of the Europeana query log. Furthermore, Section 4 presents a novel query recommendation technique based on the knowledge extracted from query logs and, finally, Section 5 presents some conclusions and outlines possible future work.

2 Related Work

Some important efforts have been spent in the past to study how people interact with IR systems[4] by analyzing the historical search data of their users [10,17,20,8].

[1] http://www.unwto.org/facts/eng/pdf/highlights/UNWTO_Highlights10_en_HR.pdf

[2] http://www.europeana.eu/portal

[3] http://www.assets4europeana.eu/

[4] The IR systems whose studies here we refer to do not directly deal with Web users.

Similarly, there have been several works about the understanding of user search behaviors on large scale IR systems, i.e., Web Search Engines (WSEs), still by analyzing the stream of past queries collected by query logs. Although the nature of query logs coming from large scale WSEs is different with respect to small scale IR systems, many of the benefits coming from the analysis of the former could also be useful for improving the latter.

Typical statistics that can be drawn from query logs are: query popularity, term popularity, average query length, distance between repetitions of queries or terms, etc. To this end, the very first contribution in analyzing a WSE query log comes from Silverstein *et al.* [18]. Here, the authors propose an exhaustive analysis by examining a large query log of the AltaVista search engine containing about a billion queries submitted in a period of 42 days by approximately 285 million users. The study shows some interesting results including the analysis of the query sessions for each user, and the correlation among the terms of the queries. Similarly to other works, authors show that the majority of the users (in this case about 85%) visit the first page of results only. They also show that 77% of the users' sessions end up just after the first query.

Lempel and Moran [12] and Fagni *et al.* [7] study the content of another publicly available AltaVista log. This log refers to the summer of 2001 and consists of 7,175,648 issued queries, i.e., about three order of magnitude less queries than the log used by Silverstein *et al.*. Furthermore, no information about the number of logged users is released however, although this second log is smaller than the first one, it still represents a good picture of search engine users.

On average, queries issued to WSEs are quite short. Indeed, the average length of a query in the 1998 Excite log is 2.35 terms. Moreover, less than 4% of the queries contains more than 6 terms. In the case of the first AltaVista log, the average query length is slightly greater: 2.55. These numbers are deeply different compared with classical IR systems where the length of a query ranges from 7 to 15 terms. A possible explanation of this phenomenon could be that the Web is a medium used by people that strongly differ from each other in terms of age, race, culture, etc. who look for disparate information. On the other hand, traditional IR systems are instead exploited by professionals and librarian, i.e., "skilled" users, which are able to look for very focused information by precisely formulating their information needs.

Moreover, a very useful information that could be extracted from query logs are *search sessions*, i.e., sets of user actions recorded in a limited period of time that hopefully refer to the same *information need*.

Several works have addressed the search session identification problem from raw streams of queries available in user logs. To this end, Silverstein *et al.* [18] firstly define a concept of *session* as follows: two consecutive queries are part of the same session if they are issued at most within a 5-minutes time window. According to this definition, they found that the average number of queries per session in the data they analyzed was 2.02. Similarly to this approach, He and Göker [9] use different timeouts to split user sessions of the Excite query log, ranging from 1 to 50 minutes.

Radlinski and Joachims [15] observe that users often perform a sequence of queries with a similar information need, and they refer to those sequences of reformulated queries as *query chains*. Their paper presents a method for automatically detecting query chains in query and click-through logs using 30 minutes threshold for determining if two consecutive queries belong to the same search session.

More recently, novel heuristics have been proposed for effectively discovering search session boundaries in query logs. Boldi *et al.* [5] introduce the *Query Flow Graph* as a model for representing data collected in WSE query logs. They exploited this model for segmenting the query stream into sets of related information-seeking queries, leveraging on an instance of the Asymmetric Traveling Salesman Problem (ATSP). Jones and Klinkner [11] argue that within a user's query stream it is possible to recognize particular hierarchical units, i.e., *search missions*, which are in turn subdivided into disjoint *search goals*. Given a manually generated ground-truth, the authors investigate how to *learn* a suitable binary classifier, which is aimed to precisely detect whether two queries belong to the same session or not.

Finally, Lucchese *et al.* [13] devise effective techniques for identifying *task-based sessions*, i.e. sets of possibly non contiguous queries issued by the user of a Web search engine for carrying out a given *task*. Furthermore, authors formally define the *Task-based Session Discovery Problem* (TSDP) as the problem of best approximating a a *ground-truth* of manually annotated tasks, and propose several variants of well-known clustering algorithms, as well as a novel efficient heuristic algorithm, specifically tuned for solving the TSDP. Results show that it performs better than state-of-the-art approaches, because it effectively takes into account the *multi-tasking* behavior of users.

3 The Europeana Query Log

A query log keeps track of historical information regarding past interactions between users and the retrieval system. It usually contains tuples $\langle q_i, u_i, t_i, V_i, C_i \rangle$ where for each submitted query q_i the following information is available: i) the anonymized identifier of the user u_i, ii) the submission timestamp t_i, iii) the set V_i of documents returned by the search engine, and iv) the set C_i of documents clicked by u_i. Therefore, a query log records both the activities conducted by users, e.g. the submitted queries, and an implicit feedback on the quality of the retrieval system, e.g. the clicks.

In this work, we consider a query log coming from Europeana portal[5], relative to the time interval ranging from August 27, 2010 to February, 24, 2011. This is a six months worth of users' interactions, resulting in 1,382,069 distinct queries issued by users from 180 countries (3,024,162 is the total number of queries). We preprocessed the entire query log in order to remove noise (e.g., stream of queries submitted by software robots instead of humans).

[5] http://www.europeana.com/portal/

It is worth noticing that 1,059,470 queries (i.e., 35% out of the total) also contain a *filter* (e.g., YEAR:1840). Users can filter results by *type*, *year* or *provider* simply by clicking on a button, so it is reasonable that they try to refine retrieved results by applying a filter, whenever they are not satisfied. Furthermore, we find that users prefer filtering results by type, i.e., images, texts, videos or sounds. Indeed, we measure that 20% of the submitted queries contains a filter by type. This is an additional proof of the skillfulness of Europeana users and their willingness to exploit non trivial search tools to find their desired contents. This also means that advanced search aids, such as query recommendation, would be surely exploited.

Similarly to Web query log analysis [18], we discuss two aspects of the analysis task: i) an analysis on the *query set* (e.g., average query length, query distribution, etc.) and ii) a higher level analysis of *search sessions*, i.e., sequences of queries issued by users for satisfying specific information needs.

3.1 Query Analysis

First we analyzed the load distribution on the Europeana portal. An interesting analysis can be done on the queries themselves. Figure 1(a) shows the frequency distribution of queries. As expected, the popularity of the queries follows a power-law distribution ($p(x) \propto kx^{-\alpha}$), where x is the popularity rank. The best fitting α parameter is $\alpha = 0.86$, which gives a hint about the skewness of the frequency distribution. The larger α the larger is the portion of the log covered by the top frequent queries. Both [14] and [2] report a much larger α value of 2.4 and 1.84 respectively from a Excite and a Yahoo! query log. Such small value of α means that the most popular queries submitted to Europeana do not account for a significantly large portion of the query log. The might be explained by looking at and comparing the main characteristics both of Europeana and Web search engines users. Indeed, since Europeana is strongly focussed on the specific context of cultural heritage, its users are likely to be more skilled and therefore they tend to use a more diverse vocabulary.

In addition, we found that the average length of queries is 1.86 terms, which is again a smaller value than the typical value observed in Web search engine logs. We can argue that the Europeana user has a more rich vocabulary, with discriminative queries made of specific terms.

Figure 1(b) shows the distribution of the queries grouped by country. France, Germany, and Italy are the three major countries accounting for about the 50% of the total traffic of queries submitted to the Europeana portal.

Figure 2(a) reports the number of queries submitted per day. We observe a periodic behavior over a week basis, with a number of peaks probably related to some Europeana dissemination or advertisement activities. For example, we observe several peaks between the 18th and the 22th November, probably due to the fact that, in those days, Europeana announced to have reached a threshold of 14 million of indexed documents[6].

[6] http://www.sofiaecho.com/2010/11/18/995971_europes-cultural-heritage-online

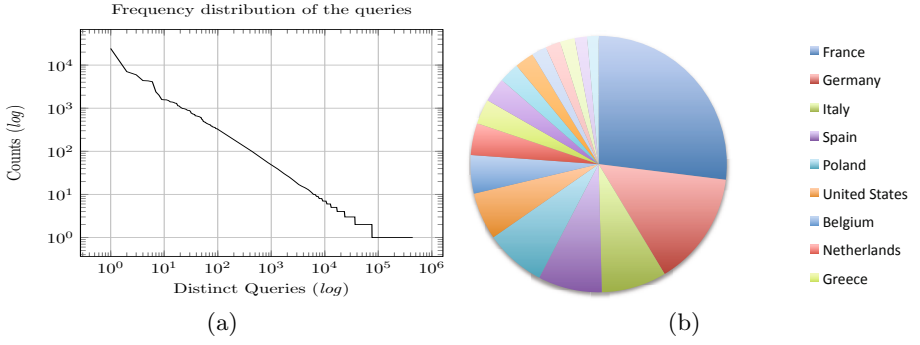


Fig. 1. Frequency distribution of queries (a) and distribution of the queries over the countries (b)

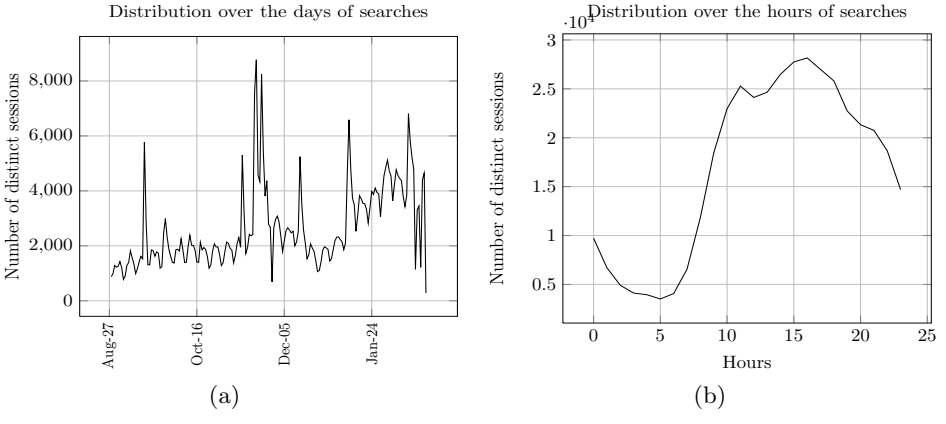


Fig. 2. Distribution of the searches over the days (a) and over the hours (b)

Figure 2(b) shows the load on the Europeana portal on a per hour basis. We observe a particular trend. The peak of load on the Europeana portal is in the afternoon, between 15 and 17. It is different from commercial Web search engines where the peak is reached in the evening, between the 19 and the 23 [4]. A possible explanation of this phenomenon could be that the Europeana portal is mainly used by people working in the field and thus, mainly accessed during working hours. From the other side, a commercial Web search engine is used by a wider range of users looking for the most disparate information needs and using it through all the day.

3.2 Session Analysis

To fully understand user behavior, it is important to analyze also the sequence of queries she submits. Indeed, every query can be considered as an improvement of the previous done by the user to better specify her information need.

Several techniques have been developed to split the queries submitted by a single user into a set of sessions [5,11,13]. We adopted a very simple approach which has proved to be fairly effective [18]. We exploit a 5 minutes inactivity time threshold in order to split the stream of queries coming from each user. We assume that if two consecutive queries coming from the same user are submitted within five minutes they belong to the same logical session, whereas if the time distance between the queries is greater, the two queries belong to two different interactions with the retrieval system.

By exploiting the above time threshold, we are able to devise 404,237 sessions in the Europeana query log. On average a session lasts about 276 sec, i.e., less than 5 minutes, meaning that, under our assumption, Europeana's users complete a search activity for satisfying an information need within 5 minutes. The average session length, i.e., the average number of queries within a session, is 7.48 queries. This number of queries is an interesting evidence that the user is engaged by the Europeana portal, and she is willing to submit many queries to find the desired result.

Moreover, we distinguish between *successful* and *unsuccessful* sessions. According to [6], a session is supposed to be successful if its *last* query has got a click associated. To this end, we find 182,280 occurrences of successful sessions in the Europeana query log, that is about 45% of the total. We notice that in [6] it was observed a much larger fraction of successful sessions, about 65%.

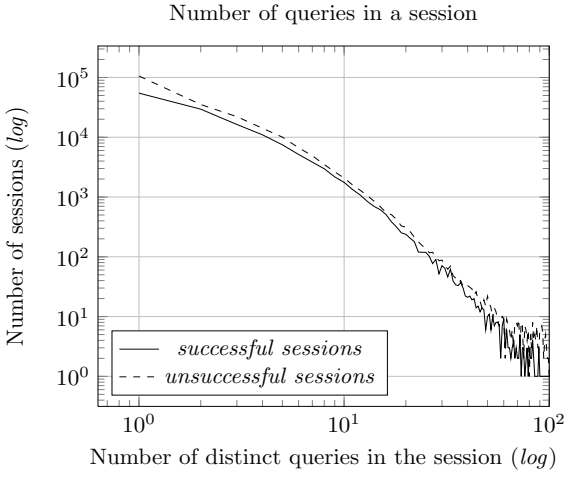


Fig. 3. Distribution of successful and unsuccessful sessions lengths (in queries)

Figure 3 shows the distributions of session lengths, both for successful and unsuccessful sessions. On the x-axis the number of queries within a session is plotted, while on the y-axis the frequencies, i.e., how many sessions to contain a specific number of queries are reported. We expect successful sessions contain

Table 1. Europeana vs. Web Search Engines: a comparison on query log statistics

	Europeana	Web Search Engines
avg. query terms	1.86	2.35 [14], 2.55 [18]
query distribution (i.e., power-law's α)	0.86	2.40 [14], 1.84 [2]
avg. queries per session	7.48	2.02 [18]
% of *successful* sessions	45	65 [6]

on average less queries than unsuccessful ones, due to the ability of the retrieval system to return early high quality results in successful session. The fact that the session length distributions are very similar, suggests that high quality results are not in the top pages, and that the Europeana ranking can be improved in order to present interesting results to the user earlier, thus reducing the successful session length with a general improvement of the user experience.

Table 1 shows some statistics extracted both from the analysis of the Europeana query log as well as from general purpose Web Search Engines historical search data.

4 A Query Recommender System for Europeana

The analysis conducted in the previous section shows that the search experience of the user interacting with Europeana could be improved. To this extent, we now introduce an application exploiting the knowledge extracted from the Europeana query log aiming at enhancing the interaction of users by suggesting a list of possible interesting queries.

A search session is an interactive process where users continuously refines their search query in order to better specify their information need. Sometimes, the successful query is not known in advance, but users might adopt concepts and terminologies also on the basis of the results pages visited. Query recommendation is a very popular technique aiming at proposing successful queries as early as possible. The approach described below, exploits successful queries from successful session to *recommend queries that allowed "similar" users, i.e., users which in the past followed a similar search process, to successfully find the information they were looking for*, and it is able to catch non trivial semantic relationships among queries.

We adopt the *Search Shortcuts* (SS) model proposed in [3] and its terminology. The SS has a clear and sound formulation as the problem of recommending queries that can reduce the search session length, i.e., leading users to relevant results as early as possible.

Let $\mathcal{U}$ be the set of users of a WSE whose activities are recorded in a query log QL, and $\mathcal{Q}$ be the set of queries in QL. We suppose QL is preprocessed by using some session splitting method (e.g. [11,13]) in order to extract query *sessions*, i.e., sequences of queries which are related to the same user search task. Formally, we denote by $\mathcal{S}$ the set of all sessions in QL, and σ^u a session issued by user u. Moreover, let us denote with σ^u_i the i-th query of σ^u. For a

session σ^u of length n its *final query* is the query σ^u_n, i.e. the last query issued by u in the session. To simplify the notation, in the following we will drop the superscript u whenever the user u is clear from the context.

As previously introduced, we say that a session σ is *successful* if and only if the user has clicked on at least one link shown in the result page returned by the WSE for the final query σ_n, *unsuccessful* otherwise.

We define a novel algorithm that aims to generate suggestions containing only those queries appearing as final in successful sessions. The goal is to suggest queries having a high potentiality of being useful for people to reach their initial goal. In our view, suggesting queries appearing as final in successful sessions is a good strategy to accomplish this task.

The SS algorithm works by efficiently computing similarities between partial user sessions (the one currently performed) and historical successful sessions recorded in a query log. Final queries of most similar successful sessions are suggested to users as search shortcuts.

Let σ' be the current session performed by the user, and let us consider the sequence τ of the concatenation of all terms with possible repetitions appearing in $\sigma'_{t|}$, i.e. the head of length t of session σ'. Then, we compute the value of a scoring function $\delta(\tau, \sigma^s)$, which for each successful session measures the similarity between its queries and the set of terms τ. Intuitively, this similarity measures how much a previously seen session overlaps with the user need expressed so far (the concatenation of terms τ serves as a bag-of-words model of user need). Sessions are ranked according to δ scores and from the subset of the top ranked sessions we suggest their final queries. It is obvious that depending on how the function δ is chosen we may have different recommendation methods. In our particular case, we opt for δ to be the similarity computed as in the BM25 metrics [16]. The choice of an IR-like metric allows us to take much care of words that are discriminant in the context of the session to which we are comparing. BM25, and other IR-related metrics, have been designed specifically to account for that property in the context of query/documents similarity. We borrow from BM25 the same attitude to adapt to this condition. The shortcuts generation problem has been, thus, reduced to the information retrieval task of finding highly similar sessions in response to a given sequence of queries. In most cases, it is enough to use only the last submitted query to propose optimal recommendations.

The idea described above is thus translated into the following process. For each unique *final query* q_f contained in successful sessions we define what we have called a *virtual document* identified by its *title* and its *content*. The title, i.e., the identifier of the document, is exactly query string q_f. The content of the virtual document is instead composed of all the terms that have appeared in queries of all the successful sessions ending with q_f. At the end of this procedure we have a set of virtual documents, one for each distinct final query occurring in some successful sessions. Just to make things more clear, let us consider a toy example. Consider the two following successful sessions: (*dante alighieri* $\rightarrow$ *divina commedia* $\rightarrow$ *paolo e francesca*), and (*divina commedia* $\rightarrow$ *inferno canto V* $\rightarrow$ *paolo e francesca*). We create the virtual document identified by title *paolo*

e francesca and whose content is the text (*dante alighieri divina commedia divina commedia inferno canto V*). As you can see the virtual document actually contains also repetitions of the same terms that are considered in the context of the BM25 metrics. All virtual documents are indexed with the preferred Information Retrieval system, and generating shortcuts for a given user session σ' is simply a matter of processing the query $\sigma'_{t|}$ over the inverted file indexing such virtual documents. We know that processing queries over inverted indexes is very fast and scalable, and these important characteristics are inherited by our query suggestion technique as well.

The other important feature of our query suggestion technique is its robustness with respect to rare and singleton queries. Singleton queries account for almost 50% of the submitted queries [19], and their presence causes the issue of the sparsity of models [1]. Since we match τ with the text obtained by concatenating all the queries in each session, we are not bound to look for previously submitted queries as in the case of other suggestion algorithms. Therefore, we can generate suggestions for rare queries of the query distribution whose terms have some context in the query log used to build the model.

5 Conclusions

In this paper we presented a part of the work carried out within the ASSETS project with the aim of improving the usability of the Europeana Portal. We place our work in the context of user-system interaction analysis for web search engines and information retrieval applications. We reused the concepts of session identification, time series analysis, query chains and task based search when analyzing the Europeana logs. To the best of our knowledge, this is first analysis of the user interaction with a cultural heritage retrieval system.

Our analysis highlights some significative differences between the Europeana query log and the historical data collected by general purpose Web Search Engine logs. In particular, we find out that both query and search session distributions show different behaviors. Such phenomenon could be explained by looking at the characteristics of Europeana users, which are typically more skilled than generic Web users and, thus, they are capable of taking advantage of the Europeana portal features to conduct more complex search sessions.

For this reason, we believe that interesting knowledge can be extracted from Europeana query log in order to build advanced assistance functionalities, such as query recommendation. In fact, we investigated the integration of a state-of-the-art algorithm into the Europeana portal. Furthermore, the specificity of the Europeana portal opens up a wide range of possible extensions to current recommendation models, taking advantage of its multi-lingual and multi-media content, and including new kinds of recommendations, e.g., recommend queries related to events or exhibitions.

As future work we intend to study how the introduction of the query recommender system changes the behavior of users interacting with the Europeana portal. Furthermore, we want to study if the sharing of the same final queries

induces a sort of "clustering" of the queries composing the successful user sessions. By studying such relation which is at the basis of our technique, we could probably find ways to improve our methodology.

Acknowledgements. This research has been funded by the EU CIP PSP-BPN ASSETS Project. Grant Agreement no. 250527.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734–749 (2005)
2. Baeza-Yates, R., Gionis, A., Junqueira, F., Murdock, V., Plachouras, V., Silvestri, F.: The impact of caching on search engines. In: Proc. SIGIR 2007, pp. 183–190. ACM, New York (2007)
3. Baraglia, R., Cacheda, F., Carneiro, V., Fernandez, D., Formoso, V., Perego, R., Silvestri, F.: Search shortcuts: a new approach to the recommendation of queries. In: Proc. RecSys 2009. ACM, New York (2009)
4. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O.: Hourly analysis of a very large topically categorized web query log. In: Proc. SIGIR 2004. ACM Press, New York (2004)
5. Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., Vigna, S.: The query-flow graph: model and applications. In: Proc. CIKM 2008. ACM, New York (2008)
6. Broccolo, D., Marcon, L., Nardini, F.M., Perego, R., Silvestri, F.: An efficient algorithm to generate search shortcuts. Tech. Rep. 2010-TR-017, CNR ISTI Pisa (2010)
7. Fagni, T., Perego, R., Silvestri, F., Orlando, S.: Boosting the performance of web search engines: Caching and prefetching query results by exploiting historical usage data. ACM Trans. Inf. Syst. 24, 51–78 (2006)
8. Gordea, S., Zanker, M.: Time filtering for better recommendations with small and sparse rating matrices. In: Benatallah, B., Casati, F., Georgakopoulos, D., Bartolini, C., Sadiq, W., Godart, C. (eds.) WISE 2007. LNCS, vol. 4831, pp. 171–183. Springer, Heidelberg (2007)
9. He, D., Göker, A.: Detecting session boundaries from web user logs. In: BCS-IRSG, pp. 57–66 (2000)
10. Hsieh-yee, L.: Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. JASIS 44, 161–174 (1993)
11. Jones, R., Klinkner, K.L.: Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In: CIKM 2008, pp. 699–708. ACM, New York (2008)
12. Lempel, R., Moran, S.: Predictive caching and prefetching of query results in search engines. In: Proc. WWW 2003, pp. 19–28. ACM, New York (2003)
13. Lucchese, C., Orlando, S., Perego, R., Silvestri, F., Tolomei, G.: Identifying task-based sessions in search engine query logs. In: Proc. WSDM 2011, pp. 277–286. ACM, New York (2011)
14. Markatos, E.P.: On caching search engine query results. In: Computer Communications, p. 2001 (2000)
15. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: Proc. KDD 2005. ACM Press, New York (2005)

16. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (2009)
17. Siegfried, S., Bates, M., Wilde, D.: A profile of end-user searching behavior by humanities scholars: The Getty Online Searching Project Report No. 2. JASIS 44(5), 273–291 (1993)
18. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum 33, 6–12 (1999)
19. Silvestri, F.: Mining query logs: Turning search usage data into knowledge. Foundations and Trends in Information Retrieval 1(1-2), 1–174 (2010)
20. Spink, A., Saracevic, T.: Interaction in information retrieval: selection and effectiveness of search terms. JASIS 48(8), 741–761 (1997)

Implementing Enhanced OAI-PMH Requirements for Europeana

Nikos Houssos[1], Kostas Stamatis[1], Vangelis Banos[2], Sarantos Kapidakis[3], Emmanouel Garoufallou[4], and Alexandros Koulouris[5]

[1] National Documentation Centre, Greece
[2]Veria Public Library, Greece
[3]Laboratory on Digital Libraries and Electronic Publishing, Department of Archive and Library Sciences, Ionian University, Greece
[4] Technological Educational Institute of Thessaloniki, Greece
[5]Technological Educational Institute of Athens, Greece
{nhoussos,kstamatis}@ekt.gr, vbanos@gmail.com, sarantos@ionio.gr, mgarou@libd.teithe.gr, akoul@teiath.gr

Abstract. Europeana has put in a stretch many known procedures in digital libraries, imposing requirements difficult to be implemented in many small institutions, often without dedicated systems support personnel. Although there are freely available open source software platforms that provide most of the commonly needed functionality such as OAI-PMH support, the migration from legacy software may not be easy, possible or desired. Furthermore, advanced requirements like selective harvesting according to complex criteria are not widely supported. To accommodate these needs and help institutions contribute their content to Europeana, we developed a series of tools. For the majority of small content providers that are running DSpace, we developed a DSpace plug-in, to convert and augment the Dublin Core metadata according to Europeana ESE requirements. For sites with different software, incompatible with OAI-PMH, we developed wrappers enabling repeatable generation and harvesting of ESE-compatible metadata via OAI-PMH. In both cases, the system is able to select and harvest only the desired metadata records, according to a variety of configuration criteria of arbitrary complexity. We applied our tools to providers with sophisticated needs, and present the benefits they achieved.

Keywords: OAI-PMH, Europeana, EuropeanaLocal, Tools, DSpace Plug-in, Interoperability, Information integration, Metadata harvesting, Europena Semantic Elements.

1 Introduction

Europeana is an evolving service, which will constitute an umbrella of European metadata from distributed cultural organisations. Europeana currently gives access to more than 14 million items representing all Member States including film material, photos, paintings, sounds, maps, manuscripts, books, newspapers and archival papers. The Europeana service [1] is designed to increase access to digital content across Europe's cultural organisations (i.e. libraries, museums, archives and audio/visual

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 396–407, 2011.

archives). This process will bring together and link up heterogeneously sourced content, which is complementary in terms of themes, location and time. Europeana's active partner network consists of 180 organisations till now.

In order to achieve these goals, the European Union launched in June 2008 the EuropeanaLocal project in the framework of the eContentPlus program. Up to June 2011, the EuropeanaLocal partners aim to make available to Europeana more than 20 million items, held across 27 countries. At the same time, they are committed to exploring and developing efficient and sustainable processes and governance procedures so that the growing numbers of regional and local institutions can easily make their content available to Europeana in the future by adopting and promoting the use of its infrastructure, tools and standards [2].

Greece is participating in EuropeanaLocal with content providers and the Hellenic Aggregator created and supported by the Veria Central Public Library (VCPL). Since March 2010, 10 content providers, from which 7 use DSpace, have followed closely the Europeana standards, thus implementing full support for Europeana Semantics Elements (ESE) and have been harvested successfully by the VCPL Aggregator (http://aggregator.libver.gr) and Europeana [3]. In March 2011, the Hellenic Aggregator provided 130.000 items to Europeana.

One of the most important aspects in the process of creating a Europeana Compliant digital repository is the support for ESE, which is virtually a new Dublin Core Profile, developed by Europeana in order to fulfill its operational requirements. Existing digital repository software in general does not support ESE by default as it is the case with Dublin Core. Nevertheless, the nature of the formats makes it feasible to alter existing software and data in order to add support for ESE. Specific information about the process can be found at the DSpace plugin for Europeana Semantic Elements webpage [4], developed by the Veria Central Public Library (VCPL) and the Hellenic National Documentation Centre (EKT).

The first step in the process is to use the Europeana XML Namespace http://europeana.eu/schemas/ese/ and augment existing systems' configuration in order to support the additional ESE elements. After implementing ESE support, the repository has to be populated with the appropriate metadata values. This task can be either performed manually through the appropriate user interface of each digital library or automatically by using special software tools developed for this purpose. It must be noted that due to the wide usage of the DSpace software internationally and in Greece, the focus has been the implementation of tools for this specific platform.

Except from DSpace and other modern digital repository platforms, there are also numerous digital libraries built with older or closed source technologies or legacy software which do not support OAI-PMH or any other form of automatic metadata exchange. In these cases, special techniques should be applied in order to extract metadata through plain HTTP requests, for example the DEiXTo tool.

DEiXTo (or ΔEiXTo) [5] is a powerful freeware web data extraction tool, based on the W3C Document Object Model (DOM), created by an independent software developer. It allows users to create highly accurate "extraction rules" (wrappers) that describe what pieces of data to scrape from a web page. When used appropriately, DEiXTo can extract meaningful metadata from web pages of non standards compliant digital content collectionsand generate appropriate Dublin Core and ESE records.

These records can be utilised by any standards compliant metadata harvester in order to be included in Europeana.

This paper analyses a toolset for data providers that mainly targets owners of small collections that are running DSpace (i.e. the DSpace plug-in, which converts and augments the DSpace metadata according to Europeana ESE requirements) as well as systems with different software, incompatible with OAI-PMH. Focus is also on the system ability to select and harvest only the desired metadata records, according to a variety of configuration criteria of arbitrary complexity that is applied in both cases.

The structure of the rest of the present text is as follows: Section 2 describes the advanced harvesting requirements addressed by our solution and the motivation based on practical needs of data providers. Section 3 presents related work and section 4 elaborates on the actual solution. Section 5 describes the application of the proposed approach in real use cases, while the last section of the article provides summary, conclusions and plans for further work.

2 The Case for Enhanced OAI-PMH Compliant Data Providers

The ubiquitous OAI-PMH protocol provides an interoperability framework based on metadata harvesting. Two types of entities exist in a typical OAI-PMH interaction: the data provider that exposes metadata to interested clients and the service provider that offers value-added services on top of metadata collected from data providers.

The recent proliferation of repositories worldwide has created a favourable environment for the emergence of content aggregators that act as OAI-PMH service providers collecting metadata-only records from individual data sources. Aggregators provide unified search and browse functionality as well as the foundation and infrastructure for advanced value-added services that become particularly meaningful when provided over content of substantial size. A number of important aggregators with international coverage and diverse scope have entered the scene in the last few years. Distinctive examples are Europeana, the European digital heritage gateway, DRIVER and OpenAIRE (repositories of peer-reviewed scientific publications) and DART Europe (European portal to research theses and dissertations).

Compatibility with aggregators is nowadays a sine qua non pre-requisite for repositories, since it provides increased visibility, enables content re-use and allows participation of individual collections to the evolving global ecosystem of interoperable digital libraries. In this context, it is becoming an increasingly common requirement for repositories to provide for retrieval by an aggregator only a subset of the metadata records it contains, essentially enabling *selective harvesting*. This may be needed for various reasons; certain indicative use cases include the following:

- The aggregator collects only records that meet specific criteria concerning IPR, copyright and open access:
  - Records are included in the harvesting set only when there is a freely accessible digital item (eg full text articles, books, etc.). Such policies are followed by Europeana, DRIVER, OpenAIRE and DART Europe.
  - Only metadata records which are themselves freely available for various uses, ideally through appropriate licensing (e.g. Creative Commons). This is required, for example, by Europeana.

- Thematic aggregators collect only records for content in specific subject areas, while individual repositories can be interdisciplinary. Such is the case with the VOA3R aggregator on Agriculture and Aquaculture. Europeana can be also considered an analogous example, since in initial stages of development concentrates on collecting mainly cultural heritage content (e.g. peer-reviewed journal articles are not included).
- The aggregator collects only records for content of a specific type (e.g. theses, like DART Europe), while individual repositories may contain different types.

The above indicate the complexity of supporting selective harvesting. This requirement becomes more difficult to achieve when you consider that a repository is likely to provide records to more than one aggregators, each with different requirements. Typically, OAI-PMH sets are implemented within repository platforms in a static fashion, through the creation of one set per individual collection in the repository. This approach is clearly not sufficient because, as is evident from the above examples, the desired sets to harvest may contain records spread over different collections. For practical needs to be satisfied and capabilities provided by the OAI-PMH sets specifications to be fully exploited, more sophisticated mechanisms are required, for example "virtual" sets that are dynamically formed per request based on specific conditions – a solution perfectly compatible with OAI-PMH.

Another important aspect and use case of selective harvesting is the retrieval of records from systems that are not compliant with OAI-PMH. These might include legacy systems like custom, non-standard databases, bibliographic catalogs of Integrated Library Systems connected with the corresponding digital material, etc. A common case is that such systems contain an array of diverse records, many of them not relevant for particular aggregators. Therefore, filtering needs to be applied, possibly according to complex criteria with a local, collection-dependent character. Crucial aspects for the success of this task are the adoption of a systematic way of implementing and injecting into the harvesting logic the filtering functionality, as well as repeatability of this procedure that enables periodic updates of metadata in the aggregator that reflect changes of records within the source systems. It is worth noting that the optimal option for content providers of this kind would be to provide their digital content through a repository platform, so that a holistic, standards-compliant solution is applied for the management of their digital material and metadata, enabling advanced services such as digital files preservation, curation, persistent identification, full-text indexing, etc.; however, this might not be feasible in the near term (e.g. due to lack of resources).

Addressing the above requirements and issues constitute the main aims of the system and approach presented in this paper, elaborated in Section 4.

3 Related Work

Mazurek et al [6], present the idea, role and benefits of a selective harvesting extension of the OAI-PMH protocol, developed and applied in Polish digital libraries in frame of the ENRICH project. Specifically, they describe the OAI-PMH protocol extension developed by the Poznan Supercomputing and Networking Center, which

allows harvesting of resources based on a search query specified in the Contextual Query Language. This selective harvesting extension is being used by the Polish national aggregator, which enables extended selective harvesting at the national level. It is notable that in this approach filtering criteria are specified directly from the side of the aggregator.

The concept, implementation and practical application of the OAI-PMH protocol extension is also presented at the Mazurek, Mielnicki and Werla [7] JCDL 2009 poster.

Finally, Sanderson Young and LeVan [8], briefly contrast the information retrieval protocols SRW/U (the Search/Retrieve Web service) and OAI (Open Archives Initiative), their aims and approaches, and then, they describe ways in which these protocols have been or may be usefully co-implemented.

A common limitation of the aforementioned approaches is that data is retrieved from data sources through queries in standard query languages like CQL. In practical situations it is frequently the case that such queries cannot fulfill the custom and complex selective harvesting requirements for data providers, as demonstrated also in the use case of paragraph 5.2. Furthermore, this solution requires a full-fledged query language to be implemented against a variety of back-end systems / data sources, while the approach proposed in this paper requires from data providers to implement only the specific bulk data loaders and filters that are necessary / useful in their particular case.

The University of Minho has developed an OAI Extended AddOn for DSpace [9], which enables selective harvesting through the incremental, piece-wise addition of objects like filters in the OAI-PMH server. The solution is bound to DSpace and does not support retrieval from legacy, non OAI-compliant sources, since, compared with our approach, there is no abstraction neither of the data records nor the data loading and output generation functionalities.

4 An Innovative Approach to Implementing Enhanced Data Providers

The main idea of our approach is to enhance an OAI-PMH server (data provider) with a number of important capabilities particularly related to selective harvesting, while maintaining full compatibility with the protocol and respecting the OAI-PMH "contract" towards clients. These capabilities are the following:

- Dynamic definition of sets and their membership, possibly based on complex criteria that do not correspond to the coarse-grained and static classification of repository records in pre-defined sets and cannot be expressed with typical query languages used by systems like federated search platforms.
- A systematic way to introduce to an OAI-PMH server implementation advanced logic necessary for selective harvesting such as transformations among different formats and schemata, filtering and updating of data. Incremental development and piece-wise enhancement of selective harvesting logic at fine levels of granularity are important relevant requirements as is the simplicity and separation of concerns among developers of different parts of the OAI-PMH data provider.

For example, the technical person creating or updating filters and crosswalks for the implementation of harvesting use cases should not need to be aware of harvesting or OAI-PMH specific technology and can thus concentrate on improving the filtering or update functionality per se.

- Support of a modular implementation that enables retrieval of metadata records from a variety of non OAI-PMH sources via simple extensions to the core architecture for data loading, tranformation and exporting in the desired formats and schemata. This is highly important, since vast sets of important content are "hidden" behind legacy, custom-made applications that do not follow state-of-t-art interoperability standards and are thus deprived of their potentially significant impact for end users and other stakeholders like value-added services developers.

To achieve the above, we have designed according to these principles and developed a modular component called transformation engine. This component has been successfully incorporated in OAI-PMH server implementations for two types of systems: (a) OAI-PMH–compliant repositories, in particular running the DSpace platform, that have been enriched with selective harvesting functionality and (b) Z39.50-compliant bibliographic catalogs of metadata records, possibly with links to digital material, that have been enhanced with OAI-PMH data providers which enable pre-processing, mapping metadata entries to OAI-PMH clients requirements and also support repeatability of the procedure at periodic time intervals, as is common for OAI-PMH compliant sources.

The rest of this chapter is structured as follows: First, a detailed description of the transformation engine is provided, followed by a report on the implementation of the two aforementioned distinct use cases.

4.1 The Transformation Engine

The transformation engine is a generic framework for implementing data transformation workflows. It allows the decoupling of communication with third party data sources and sinks (e.g. loading and exporting/exposing data) with the actual tasks

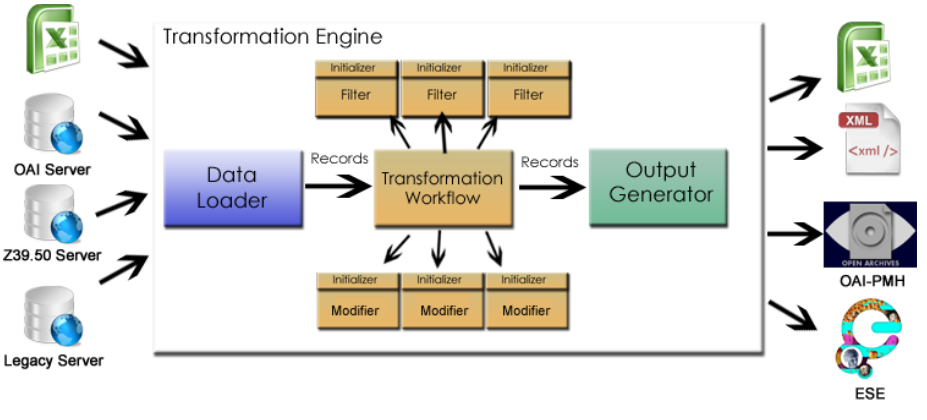


Fig. 1. Architecture of the transformation engine

that comprise the transformation. Furthermore, it enables the decomposition of a workflow into autonomous, modular pieces (transformation steps), facilitating the continuous evolution/re-definition of workflows to constantly changing data sources and the development of fine-grained workflow extensions in a systematic way. It is worth noting that the transformation engine is an independent component that is used in a modular fashion in the proposed toolset. It has been used by EKT as an autonomous module in a variety of contexts, for example for the population of digital repositories of Greek public libraries [10] with metadata from ILS catalogues.

A key aspect of the engine's design is the *Record* abstraction. Metadata records are represented by a hierarchy of classes extending the abstract Record class. A simple common interface for all types of records proved adequate to allow complex transformation functions. Examples of record implementations that have been implemented and used until now concern UNIMARC, MARC21, Dublin Core, ESE, various structured formats for references (e.g. BibTex, RIS, Endnote) while there is also a more general abstraction for XML records. The main methods of the Record interface are shown in the following:

```
public abstract List<String> getByName(String elementName);
public abstract void removeField(String fieldName);
public abstract void addField(String fieldName,
ArrayList<String> fieldValues);
public void updateField(String fieldName, ArrayList<String>
fieldValues)
```

As depicted in Figure 1, data loaders are used to read data from external sources (e.g. files, repository databases, Z39.50 servers, even OAI-PMH data providers) and forward it to the transformation workflows in the form of a certain syb-type of Record. The output generators undertake the exporting / exposing of records to third party systems and applications. The transformation workflow(s) is the place where the actual tasks are executed. A workflow consists of processing steps, each of which falls most of the time into one of the two following categories: *Filters* determine whether an input record will make it to the output. *Modifiers* can perform operations on record fields and their values (e.g. add/remove/update field). *Initializers* initialize data structures that are used by processing steps. By using the record interface in the implementation of entities like filters and modifiers a great degree of separation of concerns is achieved (for example, knowledge of the specifics of MARC is not necessary for a developer to create a modifier that performs some changes on an input MARC record).

A workflow is defined as a series of processing steps in a configuration file outside the source code of the engine, in particular using the dependency injection mechanisms of the Spring framework. Thus, a tranformation engine system can include many data loaders, output generators and transformation steps, but a specific scenario (being described a Spring configuration XML file) can make use of only some of them according to the user needs.

4.2 Extending the OAI-PMH-Compliant Harvesting Server of a Repository

An obvious use case of the proposed mechanism is the enhancement of modern repository platforms that already support OAI-PMH with the aforementioned advanced functionality. In particular, we have incorporated the transformation engine in the OAI-PMH module of the DSpace platform, which is the most popular repository platform in Greece (also among the contributors to Europeana Local).

In the vanilla DSpace platform, the harvesting server receives requests through the DSpaceOAICatalog module, where record filtering is performed, if required, according to the specifications of OAI-PMH, based on time stamps or set membership. Following this stage and before sending results to the client, the DSpaceOAIClosswalk addresses adaptation of the returned records (e.g. modification of the exposed metadata schema, appropriate adjustments in field values).

This procedure is carried out by the DSpaceOAICatalog and the DSpaceOAICrosswalk classes depicted in Figure 2.

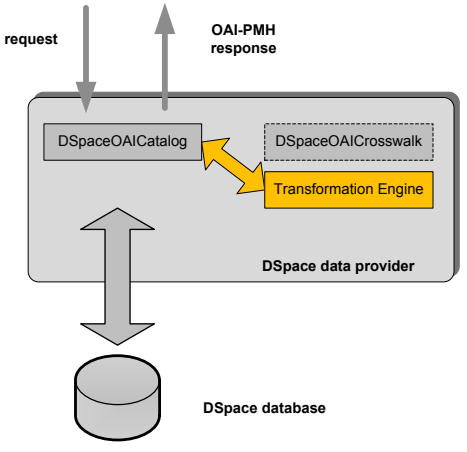


Fig. 2. Enhanced DSpace data provider

In the proposed enhanced version, the architecture of the DSpace data provider is modified as depicted in Figure 2. The tasks of record filtering and record adaptation according to the desired output schema (e.g. ESE) are handled by the Transformation Engine that is injected into the OAI-PMH server implementation, with Filters undertaking selection of records and Modifiers the work of the metadata crosswalk. Selective harvesting is based on virtual, dynamic sets. A virtual set is essentially defined as the set of repository records that results from a distinct transformation workflow, i.e. a series of specific filters and modifiers applied on repository metadata records, as specified in a Spring configuration file. If a particular record is not filtered during the workflow it is considered a member of the virtual set and is included in the record set returned to the client.

For the case of Europeana /ESE, specific user-defined classes have been developed and injected into the transformation engine (e.g. ESERecord, ESEOutputGenerator,

ESEMappingModifier) in a straightforward manner, demonstrating the ease of system customisation for developers which are due to the separation of concerns enforced by the engine's modular design.

4.3 Enabling OAI-PMH-Compliant Harvesting of MARC/Z39.50 Data Sources

Large volumes of valuable content are hosted today in systems that are not compliant with OAI-PMH and thus providing them to aggregators like Europeana is a challenging task. In this use case, based on the DSpace OAI-PMH module, we have developed an OAI-PMH server that reads UNIMARC data records from Z39.50 data sources and serves them to OAI-PMH clients (and in particular Europeana), as depicted in Figure 3. To achieve this, we modified the DSpaceOAICatalog so that upon receiving a request it triggers the transformation engine. A MARC/Z39.50 data loader is invoked first to get UNIMARC records (in ISO 2709 or MARCXML format) from a standard Z39.50 server, using the JZKit open source library, and transform them, based on the MARC4J tool, into MARCRecord objects (MARCRecord is an abstraction for MARC records following the aforementioned Record interface). These objects are relayed to the transformation workflow where filters are applied for tasks like rejection of records that do not have associated digital

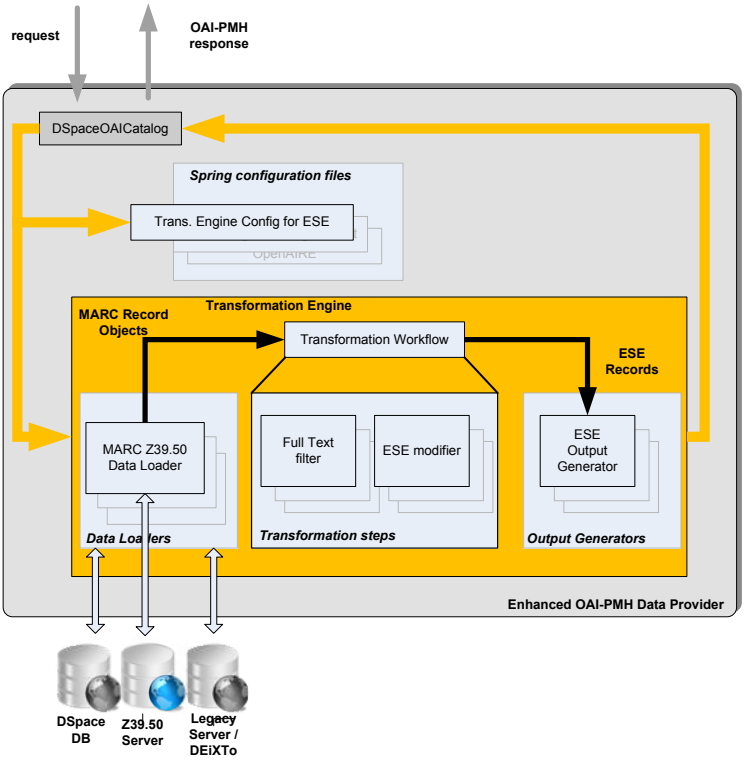


Fig. 3. Architecture for OAI-PMH compliant harvesting of non OAI-PMH compliant data sources

files (e.g. bibliographic records where full text is not available), de-duplication of records (in real-life cases, duplicate records may result from retrieval from different collections, even within the same data source) and modifiers are executed to transform records to the ESE format and perform various modifications to field values (e.g. normalisation, adjusting value encoding to Europeana standards). Finally, an ESE output generator provides the output in the format prescribed by Europeana.

Moreover, as Figure 3 depicts, the *Transformation Engine* can include a pool of data loaders, output generators and transformation steps allowing the system to use any of them for providing data to dissimilar aggregators. And this is possible due to the system configuration which can be done outside the source code, through XML configuration files. These files are responsible to initialise the Transformation Engine with a specific set of transformation steps that will finally produce the right outcome for the specific aggregator. Thus, the same engine instance can produce totally different results depending on the needs of a particular aggregator / harvesting case.

It is worth noting that this approach makes the harvesting process periodically repeatable even when the underlying data sources are not OAI-PMH compatible. Furthermore, evolution and requirement changes are easily catered for due to the fine-grained extensibility and modifiability of the transformation engine (e.g. a change in requirements can be normally easily addressed by writing new filters / modifiers and including them in the processing workflow and/or by updating existing ones, without any modification of the core system).

A similar architecture but with more complex logic for data loading and mapping needs to be applied in the case of data sources not following standard metadata schemata, for example custom databases of digital material or even unstructured information in static web pages. Addressing the latter case can be assisted by tools like DEiXTo, which has been employed also within Europeana Local for collecting metadata from Greek sources.

5 Real Use Cases

5.1 The Environment and Data Sets

The Technical Chamber of Greece wants to contribute to Europeana collections that contain all their current publishing work (TEE digital library), some historical editions (1932-1980), and their multimedia content on engineers, buildings and posters.

The descriptions of these objects are in the UNIMARC format, mixed with descriptions without online objects, which are inappropriate for Europeana. Additionally, their own content management system provides the above 5 collections together with other content, from their own regional subdivisions, their journal subscriptions, etc. The right selection or records has to be performed before they become available to Europeana.

The metadata records that could be finally contributed to Europeana are approximately 6800. The most frequent metadata field is dc:subject, which is usually repeated at least 4 times, and the 28284 subjects that appear, contain 4669 unique values. The lengthiest field is dc:title with 18 words on average and follows

dc:description and dcterms:isPartOf with 15, while the dcterms:isPartOf is used in the 97% of the records, and most fields are included once on each record.

Another case, corresponding to enhancing already OAI-PMH compatible data sources, has been the ability to provide virtual sets/collections of metadata records in the Greek National Archive of Doctoral Dissertation repository (http://www.didaktorika.gr / HEDI – a service operated by the National Documentation Centre) to harvesting clients. The respective repository contains more than 23.500 thesis records – each of them is assigned to one or more disciplines according to the Frascati classification. More than 1.000 of them belong to Agricultural Sciences class or its sub-classes and have been contributed to the VOA3R thematic aggregator (virtual repository) covering the areas of agriculture and aquaculture [11].

5.2 Two Practical Applications of the Approach

The most interesting and challenging case of application of the proposed system has been the delivery of ESE-compliant metadata from UNIMARC records in Z39.50 sources, which was done for the Technical Chamber of Greece. The retrieval of the desired sets of records was not possible using only queries (e.g. PQF or CQL) to the Z39.50 server, since the criteria for filtering where quite custom and complex, (e.g. availability of full-text that was speficied in a non-standard way in the metadata records, filtering of records that are present in the database but are not published by the Technical Chamber of Greece, etc.) and also de-duplication of records was required. Using appropriate queries our data loader retrieves an unfiltered super-set of the appropriate record set, applies the filters, applies the mapping to ESE and produces and provides to clients the metadata in ESE format. The whole procedure is repeatable and transparent to harvesting clients, which receive the ESE data through OAI-PMH without being aware of the underlying complexity. Furthermore, development of filters and modifiers does not require any knowledge of the MARC and Z39.50 standards and the structure of MARC records.

In the second case, that of VOA3R, there has been the ability to provide virtual sets/collections of metadata records in the HEDI repository to harvesting clients. One virtual set is provided for each field of science and technology as specified in the Frascati classification – a relevant field exists in each metadata record. This scheme is being used to provide metadata from this repository to the VOA3R virtual repository.

6 Summary – Conclusions and Future Work

Global efforts, like Europeana, that address many small and heterogeneous content providers, have indicated the need for advanced tools, to handle common, or less common, content provider problems. We identified several of those needs, and developed appropriate tools, to facilitate the harvesting setup and configuration.

With the proposed approach, their OAI-PMH server can apply advanced logic for selective harvesting such as transformations among different formats and schemata, filtering and updating of data. Content providers can define dynamic sets to contribute to Europeana and memberships, without altering their collections. Even when their software does not support OAI-PMH, they can use our modular implementation that enables retrieval of metadata records from a variety of non OAI-PMH sources.

We implemented these tool and extensions and used them in the context of Europeana providers, to cover their practical needs. This way, they do not have to perform such task manually, or re-implement functionality that others also implement or need, and their participation to Europeana will be easier and more flexible, according to their own collection setup and requirements.

Further work is being planned along various paths. The case studies provided clear indications that the proposed approach leads to very good performance both in terms of harvesting speed and consumption of computing and memory resources. A detailed investigation of performance issues is an interesting extension of the present work. Other plans include the incorporation of the developed modular tools into various open source OAI-PMH servers, as well as the application of the proposed approach with more content providers and a systematic user study to capture their experiences with the tools in terms of utility and ease of configuration and extension.

References

1. Koninklijke Bibliotheek: Europeana (2009), http://www.europeana.eu
2. McHenry, O.: EuropeanaLocal – its role in improving access to Europe's cultural heritage through the European Digital Library. In: 11th Annual International Conference «EVA Moscow», Moscow 2008 (2008), http://conf.cpic.ru/upload/eva2008/reports/dokladEn_1509.pdf
3. Koulouris, A., Garoufallou, E., Banos, E.: Automated metadata harvesting among Greek repositories in the framework of EuropeanaLocal: dealing with interoperability. In: Proceedings of the 2nd Qualitative and Quantitative Methods in Libraries International Conference (QQML 2010), Chania (2010)
4. Banos, E.: DSpace plugin for Europeana Semantic Elements (ESE) (2010), http://www.vbanos.gr?p=189
5. Donas, K.: DEiXTo (2010), http://www.deixto.com
6. Mazurek, C., Mielnicki, M., Parkola, T., Werla, M.: The role of selective metadata harvesting in the virtual integration of distributed digital resources. In: ENRICH Final Conference, pp. 27–31 (2009)
7. Mazurek, C., Mielnicki, M., Werla, M.: Selective harvesting of regional digital libraries and national metadata aggregators. In: 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009), New York, pp. 429–430 (2009)
8. Sanderson, R., Young, J., LeVan, R.: SRW/U with OAI: Expected and Unexpected Synergies. D-Lib Magazine 11(2) (2005), http://www.dlib.org/dlib/february05/sanderson/02sanderson.html
9. OAI Extended AddOn: University of Minho (2011), http://projecto.rcaap.pt/index.php/lang-en/consultar-recursos-de-apoio/remository?func=fileinfo&id=337
10. Digital repositories of the public libraries of Serres and Levadia, http://ebooks.serrelib.gr, http://ebooks.serrelib.gr
11. VOA3R EU project, http://www.voa3r.eu

A Survey on Web Archiving Initiatives

Daniel Gomes, João Miranda, and Miguel Costa

FCCN: Portuguese Web Archive
Av. do Brasil, 101
1700-066 Lisboa, Portugal
(daniel.gomes,joao.miranda,miguel.costa)@fccn.pt

Abstract. Web archiving has been gaining interest and recognized importance for modern societies around the world. However, for web archivists it is frequently difficult to demonstrate this fact, for instance, to funders. This study provides an updated and global overview of web archiving. The obtained results showed that the number of web archiving initiatives significantly grew after 2003 and they are concentrated on developed countries. We statistically analyzed metrics, such as, the volume of archived data, archive file formats or number of people engaged. Web archives all together must process more data than any web search engine. Considering the complexity and large amounts of data involved in web archiving, the results showed that the assigned resources are scarce. A Wikipedia page was created to complement the presented work and be collaboratively kept up-to-date by the community.

1 Introduction

The web was invented to exchange data between scientists but it quickly became a crucial mean of publication. However, the web is extremely ephemeral. Most of its information becomes unavailable and is lost forever after a short period of time. It was observed that 80% of the pages are updated or disappear after 1 year [49]. Even printed publications suffer from the effects of web data transience because they frequently cite online resources that became unavailable [52]. Besides losing important scientific and historical information, the transience of the information published on the web causes common people to lose their memories as individuals (e.g. photos shared exclusively through the web). Broken links also degrade the performance of popular web applications and services, such as shared bookmarks, search engines or social networks, leading their users to dissatisfaction.

The web needs preservation initiatives to fight ephemerality. It must be ensured that the information besides being accessible worldwide, prevails across time to transmit knowledge for future generations. Web archives are innovative systems that acquire, store and preserve information published on the web. Notably, they also contribute to preserve contents born in non-digital formats that were afterwards digitized and published online. Web archives enable numerous new use cases. Journalists can look for information to document articles, software engineers can search for documentation to fix legacy systems, webmasters can recover past versions of their site's pages or historians can analyze web pages as they do for paper documents.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 408–420, 2011.

This study presents a survey that draws a picture of worldwide initiatives to preserve information published on the web. We gathered results about 42 web archiving initiatives and analyzed metrics, such as, the volume of archived data, used formats or number of people engaged. Considering the complexity and large amounts of data involved in web archiving, the results showed that the resources being assigned are still scarce.

During our research we observed that the publicly available information about web archives is frequently obsolete or inexistent. A complementary contribution of this study was the creation of a Wikipedia page named *List of Web Archiving Initiatives*[1], so that the published information can be collaboratively kept up-to-date.

2 Related Work

The National Library of Australia maintains a page listing the 17 major archiving initiatives to preserve web heritage around the world [36]. The book *Web Archiving* discusses issues related to the preservation of the web and refers to several initiatives [28]. The Web Archiving Workshop began in 2001 and yearly presents updated work about this field [19].

The Joint Information Systems Committee (JISC) published three studies about web archiving. One addressed the legal issues relating to the archiving of Internet resources in the United Kingdom, European Union, USA and Australia, and presented recommendations about the policies that should be adopted in the UK [6]. The second study discussed the feasibility of collecting and preserving the web and presented a review about 8 web archiving initiatives [8] and the most recent one analyzed the researchers engagement with web archives [11].

Shiozaki and Eisenschitz reported on a questionnaire survey of 16 national libraries designed to clarify how they attempt to justify their web archiving activities [51]. The conclusion was that national libraries envisage that the benefits brought by their initiatives are greater than the costs and they are struggling to respond to legal risks (e.g. legislation, contracting and opt-out policies).

The International Internet Preservation Consortium (IIPC) was founded in 2003 and is composed by institutions that collaborate to preserve Internet content for future generations [14]. In 2008, the IIPC published the results of a survey conducted to derive profiles of its members. The survey addressed issues such as membership type, staff, used tools, legal issues and selection criteria.

During December 2010, in the context of the European research project Living Web Archives, the Internet Memory Foundation conducted a survey to characterize web archiving institutions and analyze the main problems of this field in Europe [22]. Statistics regarding institution type, legal context, management and archiving policies were provided.

The 18^{th} Conference of Directors of National Libraries in Asia and Oceania published a report containing the answers obtained through a questionnaire about web archiving submitted to participant countries [34]. The answers were provided as free text and do not enable a rigorous quantitative analysis. However, they provide a rich

[1] http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives

qualitative overview about web archiving in this region of the world addressing legal frameworks, main challenges to overcome, collaborations, system descriptions, policies concerning acquisition, access and preservation. In 2010, there were 6 web archiving projects.

Our study presents an updated overview about web archiving initiatives across the world. It is the most comprehensive scientific study about web archiving. The methodology adopted differs from previous work because it was designed to obtain both quantitative and qualitative results through an interactive process with the respondents[2].

3 Methodology

Initially, this research aimed to obtain answers to the following questions about each web archiving initiative:

1. What is the name of your web archive initiative (please state if you want to remain anonymous)?
2. How many people work at your web archive (in person-month)?
3. Which is the amount of data that you have archived (number of files, disk space occupied)?

During October 2010, we tried to gather this information from the official sites and published documentation but we did not succeed because the published information was frequently insufficient or obsolete. Plus, many official sites were exclusively available on the native language of the hosting country (e.g. Chinese) and automatic translation tools were insufficient to obtain the required information. We decided to contact directly the community to complement our results. The questions were sent to a web archive discussion list, published on the site of the Portuguese Web Archive and disseminated through its communication channels (Twitter, Facebook, RSS). We obtained 27 answers. Then, we sent direct e-mails to the remaining web archives referenced by the IIPC [14], National Library of Australia [36] and Web Archiving Workshops [19]. We were able to establish contact and obtain direct answers from 33 web archiving initiatives. Finally, we sent the obtained results to the respondents for validation.

The methodology used in this research enabled web archivists to openly present information about their initiatives. For some situations, we had to actively interact with the respondents to obtain the desired information. We observed that terminology and language barriers led to different interpretations of the questions by the respondents, who involuntarily provided inaccurate answers. For instance, in question 3, we assumed that each archived file was the result of a successful HTTP download (e.g. page, image or video) but some respondents interpreted it as the number of files created to store web contents in bulk (ARC files [4]). The posterior statistical analysis of the results enabled the detection of abnormal values and correction of these errors through interaction with the respondents. We believe that the adopted methodology enabled the extraction of more accurate information and valuable insights about web archiving initiatives worldwide, than a typical one-shot online survey with closed answers. However, the cost of processing the results for statistical analysis was significantly higher.

[2] We would like to express our deep gratitude to everyone who collaborated with our survey.

Table 1. List of web archives (WA). The names of the initiatives were shortened but the references contain the official ones. The description of initiatives marked with * was exclusively gathered from publicly available information.

Initiative short name	Hosting country	Creation year	Staff		Main scope of archived content
			Full-time	Part-time	
Australia's WA [37]	Australia	1996	4	4.25	National
Tasmanian WA [54]	Australia	1996	0	1	Regional
Web@rchive [1]	Austria	2008	0	2	National
DILIMAG [18]	Austria	2007	2	0	German literature magazines
Canada WA [25]	Canada	2005	0	2	National governmental
Chinese WA* [38]	China	2003	n.a.	n.a.	National
Croatian WA [30]	Croatia	2004	4	3	National
WebArchiv [45]	Czech Republic	2000	5	0	National
Netarkivet.dk [53]	Denmark	2005	0	18	National
Finnish WA [57]	Finland	2008	2	2	National
BnF [39]	France	2006	9	0	National
INA* [17]	France	2009	n.a.	n.a.	National audiovisual
Internet Memory [23]	France, Netherlands	2004	21	0	International & service provider
Baden-Württemberg [2]	Germany	2003	7.5	0	German literature
German Bundestag* [10]	Germany	2005	n.a.	n.a.	German parliament
Iceland* [31]	Iceland	2004	n.a.	n.a.	National
WA Project [33]	Japan	2004	10	2	National
OASIS [40]	Korea	2001	3	11	National
Koninklijke Bibliotheek [46]	Netherlands	2006	1	1	National
New Zealand WA [41]	New Zealand	1999	3	10	National
National Library Norway* [42]	Norway	n.a.	n.a.	n.a.	National
Portuguese WA [12]	Portugal	2007	4	1	National
WA of Čačak [50]	Serbia	2009	0	1	Regional
WA Singapore* [35]	Singapore	n.a.	n.a.	n.a.	National
Slovenian WA [16]	Slovenia	2007	1	0	National
Preservation .ES [43]	Spain	2006	2	2	National
Digital Heritage Catalonia [26]	Spain	2006	4	0	Regional
Kulturarw3* [44]	Sweden	1996	n.a.	n.a.	National
WA Switzerland [55]	Switzerland	2008	0	3	National
NTUWAS [47]	Taiwan	2007	0	3	National
WA Taiwan* [32]	Taiwan	2007	n.a.	n.a.	National
UK WA [3]	UK	2004	n.a.	0	National
UK Gov WA [56]	UK	2004	4	2	National governmental
Internet Archive [21]	USA	1996	12	0	International & service provider
Columbia University [7]	USA	2009	3	1	Thematic: human rights
North Carolina [48]	USA	2005	0	3	Regional
Latin American* [62]	USA	2005	n.a.	n.a.	International focused on Latin America
WA Pacific Islands [61]	USA	2008	0	4	International focused on Pacific Islands
Library of Congress [27]	USA	2000	6	80	National
Harvard University Library [15]	USA	2006	0	6	Institutional
California Digital Library [5]	USA	2005	4	1	International & service provider
University of Michigan [58]	USA	2000	0	2	Institutional

4 Web Archiving Initiatives

Table 1 presents the 42 web archiving initiatives identified across the world ordered alphabetically by their hosting country. Web archiving initiatives are very heterogeneous in size and scope. The WA of Čačak aims to preserve sites related to this Serbian city, while the Internet Archive has the objective of archiving the global web. The obtained results show that 80% of the archives exclusively hold content related to their hosting country, region or institution. However, initiatives hosted in the USA like the Latin American WA, Internet Archive or the WA Pacific Islands also preserve information related to foreign countries. The creation and operation of a web archive is complex and costly. The Internet Archive, Internet Memory and California Digital Library provide

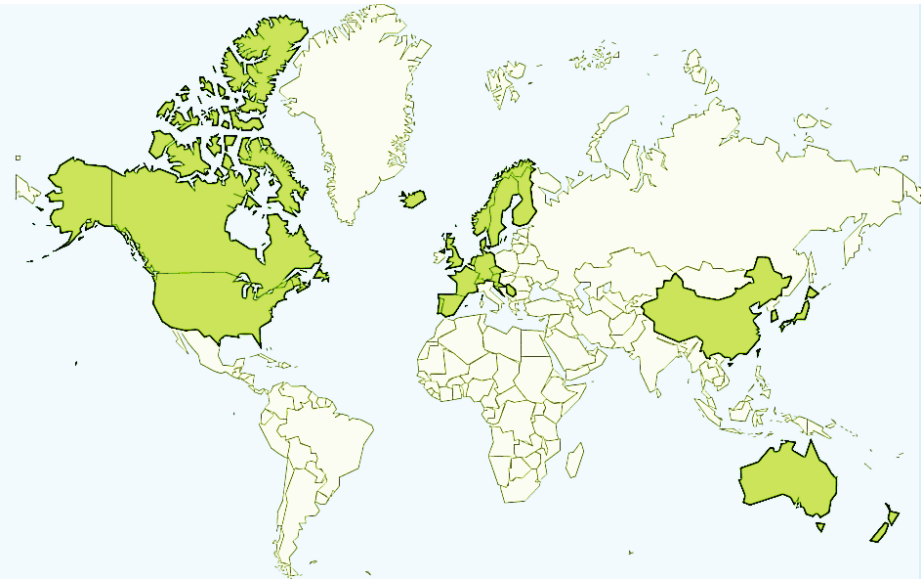


Fig. 1. Countries hosting web archiving initiatives

web archiving services that can be independently operated by third-party archivists. The services are named Archive-it[3], ArchiveTheNet[4] and Web Archiving Service[5], respectively. These services enable focused archiving of web contents by organizations, such as universities or libraries, that otherwise could not manage their own archives. For instance, the Archive-it service is used by the North Carolina, ArchiveTheNet is used by the UK Government WA and the Web Archiving Service by the University of Michigan.

The measurement of human resources engaged in web archiving activities was not straightforward (question 2). Most respondents could not provide an effort measurement in person-month. The presented reasons were that the teams were too variable and some services were hired to third-party organizations out of their control. Instead, most of the respondents described their staff and hiring conditions. The obtained results show that web archiving engages at least 112 people in full-time and 166 in part-time. The total of 277 people that preserve and provide access to the past of the web since its inception contrasts with the resources invested to provide access to a snapshot of the current web. For instance, Google by itself has 24 400 full-time employees, from which 9 508 work in research and development and 2 768 in operations [60]. The web archive teams are typically small, presenting a median staff of 2.5 people in full-time (average of 3.5) and 2 people in part-time (average of 5) and are mostly composed by librarians and information technology engineers. The results show that 11 initiatives (26%) don't have any person dedicated full-time. The effort of part-time workers is variable, for instance,

[3] http://www.archive-it.org
[4] http://archivethe.net
[5] http://webarchives.cdlib.org

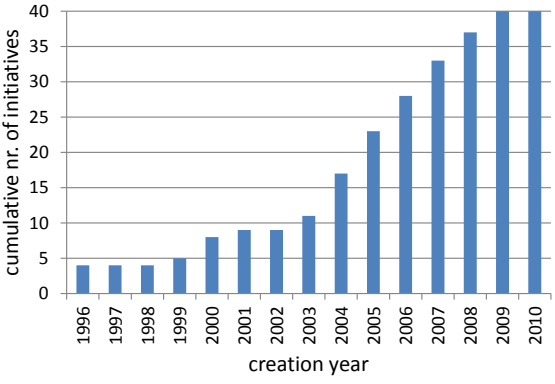


Fig. 2. Cumulative number of initiatives created per year

at the Library of Congress they spend only a few hours a month. Most of the human resources are invested on data acquisition and quality control.

Figure 1 presents the location of countries that host web archiving initiatives. The 42 initiatives are spread across 26 countries. There are 23 initiatives hosted in Europe, 10 in North America, 6 in Asia and 3 in Oceania. Half of the initiatives are hosted in countries belonging to the Organisation for Economic Co-operation and Development (OECD). From the 34 countries that belong to the OECD, 21 (62%) host at least one web archiving initiative, which is an indicator of the importance of web archiving in developed countries. Most of the countries host one (74%) or two initiatives (22%). The only country that hosts more is the USA with 8 initiatives. Although being part of a country, initiatives like the Tasmanian WA (Australia), North Carolina (USA) or Digital Heritage Catalonia (Spain) are hosted at autonomous states and aim at preserving regional content.

Figure 2 presents the evolution of the number of web archiving initiatives created per year. The first web archive named Internet Archive was founded by Brewster Kale in 1996. Three initiatives followed in 1996: the Australia's Web Archive and the Tasmanian's Web Archive from Australia, and Kulturarw3 from Sweden. Only 5 new initiatives arose during the following 6 years. However, since 2003 there was a significant and constant growth with the creation of 31 initiatives, reaching 6 initiatives per year in 2004 and 2005. One possible explanation for this fact was the concern raised by the United Nations Educational, Scientific and Cultural Organization (UNESCO) regarding the preservation of the digital heritage [59].

5 Archived Data

All web archives select specific sites for archiving. This selection is determined by factors such as consent by the authors or relevance for inclusion in thematic collections (e.g. elections or natural disasters). Eleven initiatives (26%) also perform broad crawls of the web, including all the sites hosted under a given domain name or geographical location.

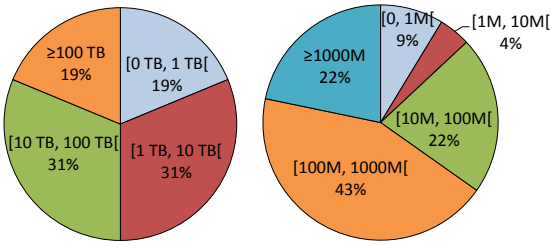


Fig. 3. Size of archived collections in: (a) Volume of data (Terabytes) (b) Number of contents (e.g. images, pages, videos)

Figure 3 presents the distribution of the archived collections measured in total volume of data and number of contents. For instance, one HTML page containing three embedded images results in the archive of four contents. The objective of this measurement was to characterize web archives regarding the total amount of data they held. Selective web archiving is frequently focused on preserving individual sites. Thus, the number of archived sites could also be an interesting metric. However, the size of web sites significantly varies and the number of archived sites by itself is not descriptive of the volume of archived data. Therefore, we decided not to include this metric to simplify the questionnaire. The results show that 50% of the collections are smaller than 10 TB and are composed by less than 1 000 million contents (78%). The volume of data correspondent to the creation of replicas to ensure preservation was not considered in this measurement. The average content size was 46 KB and ranged between 14.2 KB and 119.4 KB. There are several reasons for this difference. Some web archives are focused on specific contents which are typically large, such as video, PDF documents or images. Web archives use different formats for archiving web data that may contain additional meta-data or use compression. Another reason is that the size of contents tends to grow [29]. Therefore, older archived contents tend to be smaller than recent ones. Web archives worldwide preserved since 1996 a total of 181 978 million contents (6.6 PB). The Internet Archive by itself holds 150 000 million contents (5.5 PB). The size of the current web cannot be accurately determined. However, in 2008 Google announced that one single snapshot of the web comprised 1 trillion unique URLs (10^{12}) [13]. Notice that this number refers only to web pages and does not include contents, such as images or videos, that are also addressed by web archives. The obtained results show that the amount of archived data is small in comparison with the volume of data that is permanently being published on the web.

Figure 4 presents the distribution of the file formats used to store archived content. The ARC format was defined by the Internet Archive and applied as a *de facto* standard [4]. In 2009, the WARC format was published by the Internet Organization for Standardization (ISO) as the official standard format for archiving web contents [24] and it is already exclusively used by 10% of the initiatives. The ARC and WARC formats are dominant, being used by 54% of the initiatives. The usage of standard formats for web archiving facilitates the collaborative creation of tools, such as search engines or

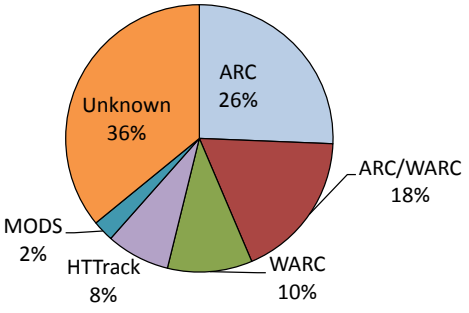


Fig. 4. Usage of file formats to store web contents

replication mechanisms, to process the archived data. Besides historical reasons, the widespread of the ARC/WARC formats was motivated by the creation of the Archive-Access project that freely provides open-source tools to process this type of files [20].

6 Access and Technologies

Figure 5(a) presents the types of search provided by the initiatives over their collections. The obtained results show that 89% of the initiatives support access to the history of a given URL, 79% enable searching meta-data and 67% provide full-text search over archived contents. There are 21 initiatives (50%) that provide full online access to search mechanisms and archived content. Some initiatives hold the copyright of the archived contents (e.g. German Bundestag, UK WA, Canada WA) or explicitly require the consent of the authors before archiving (UK WA, OASIS). The Tasmanian WA operated since its inception under the assumption that web sites fall within the definition of book. Thus, no permission to capture from publishers is required. The Internet Archive and the Portuguese WA proactively archive and provide access to contents but remove access on-demand. On the other hand, for 16 initiatives (38%) the access to the collections is somehow restricted. The Library of Congress, WebArchiv and Australia's WA provide public online access to part of their collections. Netarkivet.dk provides online access on-demand only for research purposes. The Finnish WA provides online access to meta-data but not to archived contents. BnF, Web@rchive and Preservation .ES grant access exclusively through special rooms on their facilities. Maintaining the accessibility level of the original information is mandatory to make web archives useful for citizens. If a content is publicly available on the current web, it should continue to be publicly available when it becomes a historical content. However, this policy collides with national legislations that restrict access or even inhibit proactive web archiving. The web broke economical and geographical barriers to information but legislations are raising them against historical content. It is economical unattainable for most people to travel, possibly to a foreign country, to investigate if an information published in the past exists in a web archive.

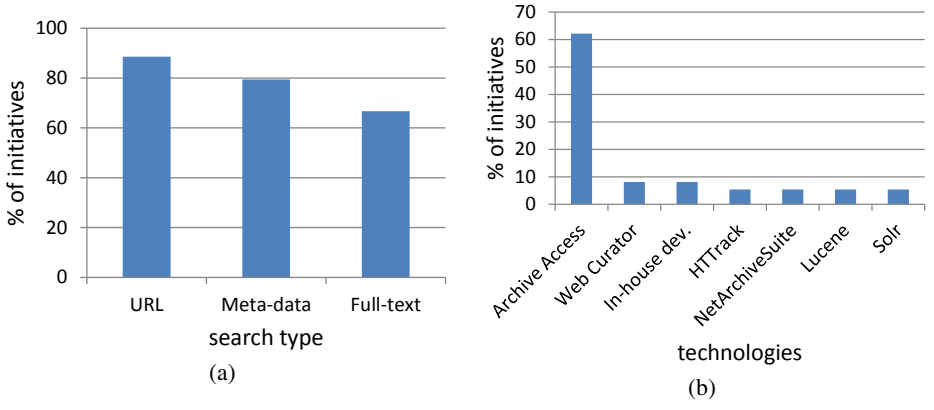


Fig. 5. Provided access to archives: (a) Access type (b) Used technologies

Figure 5(b) depicts the technologies being used by the initiatives that manage their own systems. Notice that 16% of the initiatives use software as a service to manage their collections. The Archive-Access tools are dominant (62%), including the Heritrix, NutchWAX and Wayback projects, that support content harvesting, full-text and URL search, respectively. However, respondents frequently mentioned that full-text search was hard to implement and that the performance of NutchWAX was unsatisfactory, being one reason for the partial indexing of their collections. Nonetheless, NutchWAX supports full-text search for the Finnish WA (148 million), Canada WA (170 million), Digital Heritage of Catalonia (200 million), California Digital Library (216 million) and BnF (estimated 2 100 million). Australia's WA supports full-text search over 3 100 million contents indexed using an in-house developed system named Trove. It was estimated that the largest web search engine is Google and that it indexes 38 000 million pages [9]. Creating a search engine over the archived so far (181 978 million contents), would imply indexing 4.7 times more data.

7 Conclusions

The preservation of digital heritage is crucial to modern societies because web publications are extremely transient. This study identified 42 web archiving initiatives created around the world since 1996. Web archives are typically hosted on developed countries and are composed by small teams that mainly work on the acquisition and curation of data. Most of the initiatives carefully select contents from the web to be archived. There are 3 organizations that provide web archiving services. The total amount of archived data so far reaches 6.6 PB (181 978 million contents). However, efficient search mechanisms are required to enable access to this information, which raises new technological challenges. The largest web search engine indexes only 20% of this amount of data. An additional problem are the legal barriers that restrict access to historical web

contents and diminish the visibility and importance of web archives to modern societies. Open access to historical web data would enable the creation of federated search mechanisms across web archives and the development of new applications by third-parties that would contribute to explore the potential of this valuable source of historical information. New laws regarding digital preservation and extension of the legal deposit to web contents have been approved. As future work, we intend to analyze the current legal situation worldwide regarding web archiving and its impact on cultural heritage.

Despite the social and economic impact of losing the information that is being permanently and exclusively published on the web, the obtained results show that the growing resources invested in web archiving are still relatively scarce. This fact will probably originate a historical void regarding our current times.

References

1. Austrian National Library. Österreichische Nationalbibliothek - Web archiving (March 2011), http://www.onb.ac.at/ev/about/webarchive.htm
2. Bibliotheksservice-Zentrum Baden-Württemberg. Willkommen im Bibliotheksservice-Zentrum Baden-Württemberg (March 2011), http://www.bsz-bw.de/index.html
3. British Library. UK Web Archive (March 2011), http://www.webarchive.org.uk/ukwa/
4. Burner, M., Kahle, B.: WWW Archive File Format Specification (September 1996), http://pages.alexa.com/company/arcformat.html
5. California Digital Library. Web Archives: yesterday's web; today's archives (March 2011), http://webarchives.cdlib.org/
6. Charlesworth, A.: Legal issues relating to the archiving of Internet resources in the UK, EU, USA and Australia (2003), http://www.jisc.ac.uk/media/documents/programmes/preservation/archiving_legal.pdf
7. Columbia University Libraries. Web Resources Collection Program (March 2011), https://www1.columbia.edu/sec/cu/libraries/bts/web_resource_collection/
8. Day, M.: Collecting and preserving the World Wide Web (2003), http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf
9. de Kunder, M.: WorldWideWebSize.com | The size of the World Wide Web (March 2011), http://www.worldwidewebsize.com/
10. Bundestag, D.: Web-Archiv (March 2011), http://webarchiv.bundestag.de/cgi/kurz.php
11. Dougherty, M., Meyer, E., Madsen, C., Van den Heuvel, C., Thomas, A., Wyatt, S.: Researcher engagement with web archives: State of the art. Technical report, Joint Information Systems Committee, JISC (2010)
12. FCCN. Portuguese Web Archive: search the past (March 2011), http://www.archive.pt/
13. Google Inc. Official Google Blog: We knew the web was big... (July 2008), http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
14. Grotke, A.: IIPC - 2008 Member Profile Survey Results (December 2008), http://www.netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf
15. Harvard University Library. Web Archive Collection Service - Harvard University Library (March 2011), http://wax.lib.harvard.edu/collections/home.do

16. Historical Archives of Ljubljana. Zgodovinski arhiv Ljubljana (March 2011), http://www.zal-lj.si/
17. Ina. Ina.fr - A la une: vidéo, radio, audio et publicité - Actualités, archives du jour de la radio et de la télévision en ligne (March 2011), http://www.ina.fr/
18. Innsbruck Newspaper Archive at the Univ. of Innsbruck and Dept. for Digitisation & Digital Preservation at the Univ. of Innsbruck Lib. Digitale Literatur Magazine (March 2011), http://dilimag.literature.at/default.alo
19. International Web Archiving Workshop. Index (March 2011), http://iwaw.europarchive.org/
20. Internet Archive. Nutchwax - Home Page (March 2008), http://archive-access.sourceforge.net/
21. Internet Archive. Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine (March 2011), http://www.archive.org/
22. Internet Memory Foundation. Web Archiving in Europe (2010), http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf
23. Internet Memory Foundation. Welcome to Internet Memory Foundation website (March 2011), http://internetmemory.org/en/
24. I.ISO. 28500: 2009 Information and documentation-WARC file format (2009)
25. Library and Archives Canada. Home - Library and Archives Canada (March 2011), http://www.collectionscanada.gc.ca/index-e.html
26. Library of Catalonia. PADICAT, Patrimoni Digital de Catalunya (March 2011), http://www.padicat.cat/
27. Library of Congress. Web Archiving (Library of Congress) (March 2011), http://www.loc.gov/webarchiving/
28. Masanès, J.: Web Archiving. Springer-Verlag New York, Inc., Secaucus (2006)
29. Miranda, J., Gomes, D.: Trends in Web characteristics. In: 7th Latin American Web Congress (LA-Web 2009), Merida, Mexico (November 2009)
30. National and University Library in Zagreb. Hrvatski arhiv weba, HAW (March 2011), http://haw.nsk.hr/
31. National and University Library of Iceland. Vefsafn - English (March 2011), http://vefsafn.is/index.php?page=english
32. National Central Library, Taiwan. Web Archive Taiwan (March 2011), http://webarchive.ncl.edu.tw/nclwa98Front/
33. National Diet Library. Web Archiving Project (March 2011), http://warp.da.ndl.go.jp/search/
34. National Diet Library, Japan - Conference of Directors of National Libraries in Asia and Oceania 2010. Report on questionnaire survey on web-archiving - Document 3 (2010), http://www.ndl.go.jp/en/cdnlao/meetings/pdf/report_Japan1_doc3.pdf
35. National Library Board Singapore. Web Archive - National Library Board, Singapore (March 2011), http://was.nl.sg/
36. National Library of Australia. PADI - Preserving Access to Digital Information (March 2011), http://www.nla.gov.au/padi/
37. National Library of Australia. Pandora Archive - Preserving and Accessing Networked DOcumentary Resources of Australia (March 2011), http://pandora.nla.gov.au/
38. National Library of China. Web Information Collection and Preservation - WICP (Chinese Web Archive) (March 2011), http://210.82.118.162:9090/webarchive

39. National Library of France. BnF - Digital legal deposit (March 2011), http://www.bnf.fr/en/professionals/digital_legal_deposit.html
40. National Library of Korea. About OASIS - About OASIS (March 2011), http://www.oasis.go.kr/intro_new/intro_overview_e.jsp
41. National Library of New Zealand. New Zealand Web Archive - National Library of New Zealand (March 2011), http://www.natlib.govt.nz/collections/a-z-of-all-collections/nz-web-archive
42. National Library of Norway. Nasjonalbiblioteket || index (March 2011), http://www.nb.no/
43. National Library of Spain. Biblioteca Nacional de España. Ministerio de Cultura (March 2011), http://www.bne.es/es/LaBNE/PreservacionDominioES/
44. National Library of Sweden. Swedish Websites - Kungliga biblioteket (March 2011), http://www.kb.se/english/find/internet/websites/
45. National Library of the Czech Republic. WebArchiv (March 2011), http://en.webarchiv.cz/
46. National library of the Netherlands. Web Archiving (March 2011), http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html
47. National Taiwan University Library. NTU Web Archiving System, NTUWAS (March 2011), http://webarchive.lib.ntu.edu.tw/eng/default.asp
48. North Carolina State Archives and State Library of North Carolina. North Carolina State Government Web Site Archives (March 2011), http://webarchives.ncdcr.gov/
49. Ntoulas, A., Cho, J., Olston, C.: What's new on the web?: the evolution of the web from a search engine perspective. In: Proceedings of the 13th International Conference on World Wide Web, pp. 1–12. ACM Press, New York (2004)
50. Public Library Čačak. Web Archive of Cacak - English - Digitalizacija i digitalne biblioteke (March 2011), http://digital.cacak-dis.rs/english/web-archive-of-cacak/
51. Shiozaki, R.: Role and justification of web archiving by national libraries - A questionnaire survey (2009), http://lis.sagepub.com/content/41/2/90
52. Spinellis, D.: The decay and failures of web references. Communications of the ACM 46(1), 71–77 (2003)
53. State and University Library. netarkivet.dk (March 2011), http://netarkivet.dk/index-da.php
54. State Library of Tasmania. Our Digital Island (March 2011), http://odi.statelibrary.tas.gov.au/
55. Swiss National Library. Swiss National Library NL -e-Helvetica (March 2011), http://www.nb.admin.ch/nb_professionnel/01693/index.html?lang=en
56. The National Archives. UK Government Web Archive | The National Archives (March 2011), http://www.nationalarchives.gov.uk/webarchive/
57. The National Library of Finland. Finnish Web Archive (March 2011), http://verkkoarkisto.kansalliskirjasto.fi/
58. The Regents of the University of Michigan. University of Michigan Web Archives (March 2011), http://bentley.umich.edu/uarphome/webarchives/webarchive.php
59. UNESCO. Charter on the Preservation of Digital Heritage. In: Adopted at the 32nd session of the General Conference of UNESCO (October 17, 2003), http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf
60. United States Securities and Exchange Commission. Form 10-K (December 2010), http://www.sec.gov/Archives/edgar/data/1288776/000119312511032930/d10k.htm

61. University of Hawaii at Manoa Library. Web Archiving Project for the Pacific Islands | University of Hawaii at Manoa Library (March 2011), http://library.manoa.hawaii.edu/research/archiveit/
62. University of Texas at Austin. Latin American Web Archiving Project, LAWAP (March 2011), http://lanic.utexas.edu/project/archives/

Coherence-Oriented Crawling and Navigation Using Patterns for Web Archives*

Myriam Ben Saad, Zeynep Pehlivan, and Stéphane Gançarski

LIP6, University P. and M. Curie,
4 place Jussieu 75005, Paris, France
{myriam.ben-saad,zeynep.pehlivan,stephane.gancarski}@lip6.fr

Abstract. We point out, in this paper, the issue of improving the coherence of web archives under limited resources (*e.g.* bandwidth, storage space, etc.). Coherence measures how much a collection of archived pages versions reflects the real state (or the snapshot) of a set of related web pages at different points in time. An ideal approach to preserve the coherence of archives is to prevent pages content from changing during the crawl of a complete collection. However, this is practically infeasible because web sites are autonomous and dynamic. We propose two solutions: *a priori* and *a posteriori*. As *a priori* solution, our idea is to crawl sites during the *off-peak* hours (*i.e.* the periods of time where very little changes is expected on the pages) based on patterns. A pattern models the behavior of the importance of pages changes during a period of time. As an *a posteriori* solution, based on the same patterns, we introduce a novel navigation approach that enables users to browse the most coherent page versions at a given query time.

Keywords: Web Archiving, Data Quality, Pattern, Navigation.

1 Motivation

The major challenge of web archiving institutes (Internet Archive, etc.) is to collect, preserve and enable future generations to browse off-line a rich part of the Web even after it is no more reachable on-line. However, maintaining a good quality of archives is not an easy task because the web is evolving over time and allocated resources are usually limited (*e.g.* bandwidth, storage space, etc.). In this paper, we focus on the coherence that measures how much a collection of archived pages versions reflects the real web at different points in time. When users navigate through the archive, they may want to browse a collection of related pages instead of individual pages. This collection is a set of linked pages which may or not share the same context, topic, domain name, etc. It can be a web site generally including a home page and located on the same server. But it can be also a set of interconnected web

* This research is supported by the French National Research Agency ANR in the CARTEC Project (ANR-07-MDCO-016).

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 421–433, 2011.

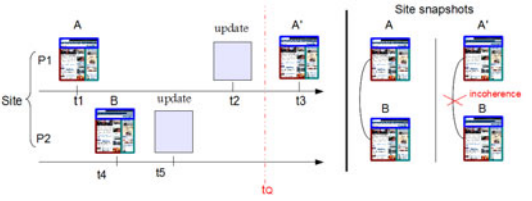


Fig. 1. Archive Coherence

pages belonging to different sites. Coherence ensures that if users reach a page version, they can also reach to the versions of other pages of the same collection, corresponding to the same point in time. In fact, during the navigation in the archive, users may browse a page version which refers to another page, but the two page versions have never appeared at the same time on the real web. This may lead to conflicts or inconsistencies between page versions, and such pages versions are considered temporally incoherent. Figure 1 depicts an example of incoherence while brwosing at time t_q.We consider two pages P_1 and P_2 of a site, updated respectively at time t_2 and t_5. A and A' are the two versions of the page P_1 captured at time t_1 and t_3. B is the only version of the page P_2 captured at t_4. If the archive is queried at time t_q, we can obtain as result the two versions A' and B because they are the "closest" from t_q. These two versions are not coherent because they have never appeared at the same time on the site (caused by pages update). However, the two versions A and B are coherent because they appear "as of" time point t_1.

The problem of incoherence usually happens when pages change their content during the crawl of an entire collection. This problem can be handled by two solutions: *a priori* and *a posteriori*. The *a priori* solution aims to minimize pages incoherence at the crawling time by adjusting crawlers strategy. The *a posteriori* solution operates at the browsing time by enabling users to navigate through the most coherent pages versions. The *a priori* solution adjusts crawlers strategy to maximize the coherence of collected pages versions independently of other quality measures (*e.g.* completeness, freshness, etc.). As it is impossible to obtain 100% of coherence in the archive due to the limited resources, an *a posteriori* solution is also needed. Thus, our challenge is to optimize browsing to enable users to navigate through the most coherent page versions at a given query time. In [2], we have discovered periodic patterns from TV channels pages which describe the behavior of (regular) changes over time. By exploiting these patterns, we propose, in this paper, novel coherence-oriented approaches of crawling and browsing to improve archives quality and users navigation.

This paper is structured as follows. In Section 2, related works are discussed. Section 3 defines a coherence measure. Section 4 describes web archiving model based on pattern. Section 5 proposes a crawling strategy to maximize archives coherence. Section 6 introduces a coherence-oriented approach to improve archive navigation. Section 7 presents preliminary results. Section 8 concludes.

2 Related Works

In recent years, there has been an increasing interest in improving coherence of web archives. In [13], authors propose a crawling strategy to improve coherence of crawled sites. However, they do not mention in which order sites should be visited. In [14], they present visualization strategies to help archivists to understand the nature of coherence defects in the archive. In another study, they define two quality measures (blur and sharp) and propose a framework, coined SHARC, to optimize pages captures. The two policies [13,9] are based on multiple revisits of web pages. However, in our work, we assume that web crawlers have limited resources which prevent from revisiting pages too often. Other studies are also closely related to our work in the sense that they aim at optimizing crawlers. To guess at which frequency each page should be visited, crawl policies are based on three major factors: (i) the relevance /importance of pages (e.g Page rank) [7], (ii) information longevity [11] and (iii) frequency of changes [5,9]. A factor that has been ignored so far is the importance of changes between pages versions. Moreover, the frequency of changes used by most policies is estimated based on homogenous poisson process which is not valid when pages are updated frequently as demonstrated in [3]. Our research is applied on the archive of French National Institute (INA) which preserves national radio and TV channels pages. These pages are updated several times a day and, hence, the poisson model can not be used as explained above. In [2], we discovered periodic patterns from TV channels pages by using statistical analysis technique. Based on patterns, we propose, in this paper, a crawl policy to improve the coherence of archives.

This paper also presents a new navigation method that takes into account the temporal coherence between the source page and the destination page. Although there are several browsers proposed to navigate over historical web data [10,15], they are only interested in navigation between versions of the same pages by showing the changes over versions. As far as we know,no approach proposes to improve the navigation in web archives by taking into account temporal coherence. The reason, as explained in [4], can be that temporal coherence only impacts the very regular users who spend lots of time navigating in the web archives. Even though, today the archive initiatives do not have many users, we believe that, popular web archives (e.g Internet Archive, Google News Archive) will get the attention of more and more regular users over web archives.

3 Coherence Measure

We define in this section a quality measure inspired by [13] which assesses the coherence of archives. The following notations are used in the paper.

- S_i is a collection of linked web pages P_i^j.
- A_{S_i} is a (historical) archive of S_i.
- $P_i^j[t]$ is a version of a page P_i^j ($P_i^j \in$ collection S_i) captured at time t.

- $Q(t_q, A_{S_i})$ is a query which asks for the closest versions (or snapshot) of A_{S_i} to the time t_q.
- $R(Q(t_q, A_{S_i}))$ is a set of versioned pages obtained as a result of querying A_{S_i} at time t_q. A' and B in Figure 1 both belong to $R(Q(t_q, A_S))$.
- $\omega(P_i^j[t])$ is the importance of the version $P_i^j[t]$. It depends on (i) the weight of the page P_i^j (*e.g.* PageRank) and on (ii) the importance of changes between $P_i^j[t]$ and its last archived version. The importance of changes between two pages versions can be evaluated based the estimator proposed in [1].

Definition 1. *Coherent Versions*

The N_i versions of $R(Q(t_q, A_{S_i}))$ are coherent, if there is a time point (or an interval) called $t_{coherence}$, so that it exists a non-empty intersection among the invariance interval $[\mu_j, \mu_{j^*}]$ of all versions.

$$\forall P_i^j[t] \in R(Q(t_q, A_{S_i})), \exists t_{coherence} : t_{coherence} \in \bigcap_{j=1}^{N_i} [\mu_j, \mu_{j^*}] \neq \varnothing \qquad (1)$$

where μ_j and μ_j^* are respectively the time points of the previous and the next changes following the capture of the version $P_i^j[t]$.

As shown in Figure 2, the three versions $P_i^1[t_1]$, $P_i^2[t_2]$ and $P_i^3[t_3]$ are coherent because there is an interval $t_{coherence}$ that satisfies the coherence constraint (1). However, the three page versions at the right are not coherent because there is no point in time satisfying the coherence constraint (1).

Definition 2. *Query-Result Coherence*

The coherence of the query result $R(Q(t_q, A_{S_i})$, also called weighted coherence, is the weight of the largest number of coherent versions divided by the total weight of the N_i versions of $R(Q(t_q, A_{S_i}))$. We assume that $\{P_i^1[t_1]$,..., $P_i^\rho[t_\rho]\}$ $\in R(Q(t_q, A_{S_i}))$ are the ρ coherent versions, i.e satisfying the constraint (1). ρ is the largest number of coherent versions composing $R(Q(t_q, A_{S_i}))$.

The coherence of $R(Q(t_q, A_{S_i}))$ is

$$Coherence(R(Q(t_q, A_{S_i})) = \frac{\sum_{k=1}^{\rho} \omega(P_i^k[t_k])}{\sum_{k=1}^{N_i} \omega(P_i^k[t_k])}$$

where $\omega(P_i^k[t_k])$ is the importance of the version $P_i^k[t_k]$.

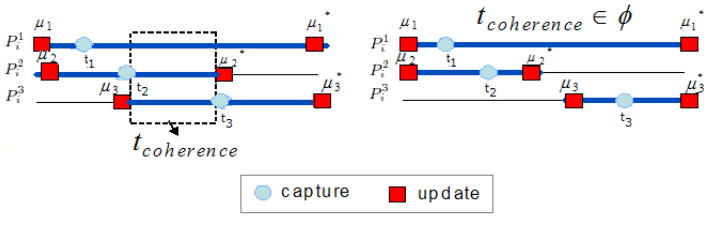


Fig. 2. Coherence Example [13]

Instead of evaluating the coherence of all the versions composing the query result $R(Q(t_q, A_{S_i}))$, we can restrict the coherence measure to only the η-top pages of A_{S_i} which are the most relevant ones. Such measure is useful to preserve particularly the coherence of the most browsed (important) pages like home pages and their related pages. Coherence of rarely browsed pages can be considered less important.

4 Pattern Model

Our work aims at improving the coherence of archived web collections by using patterns. We describe here the pattern model.

4.1 Pattern

A pattern models the behavior of page's changes over periods of time, during for example a day. It is periodic and may depend on the day of the week and of the hour within a day. Pages with similar changes behavior can be grouped to share a common pattern.

Definition 3. *Pattern A pattern of a page* P_i^j *with an interval length* l *is a nonempty sequence* $Patt(P_i^j) = \{(\omega_1,T_1); ...; (\omega_k,T_k);...; (\omega_{N_T},T_{N_T})\}$, *where* N_T *is the total number of periods in the pattern and* ω_k *is the estimated importance of changes in the period* T_k.

4.2 Pattern-Based Archiving

As shown in Figure 3, patterns are discovered from archived page versions by using an analyzer. The first step of the analyzer consists on segmenting each captured pages into blocks that describe the hierarchical structure of the page. Then, successive versions of a same page are compared to detect *structural*[1] and *content*[2] changes by using Vi-DIFF algorithm [12]. Afterwards, the importance of changes between two successive versions is evaluated based on the estimator proposed in [1]. This estimator returns a normalized value between 0 and 1. An importance value near one (respectively near 0) denotes that changes between versions are very important (respectively irrelevant *e.g.* advertisements or decoration). After that, a periodic pattern which models changes importance behavior is discovered for each page based on statistical analysis. In [2], we have presented, through a case study, steps and algorithms used to discover patterns from French TV channels pages. Discovered patterns are periodically updated to always reflect the current behavior. They can be used to improve the coherence of archives. Also, they can be exploited by the browser to enable users to navigate through the most coherent page versions as shown in Figure 3.

[1] The changes that affect the structure of blocks composing the page.
[2] The changes that modify links, images and texts inside blocks of the pages.

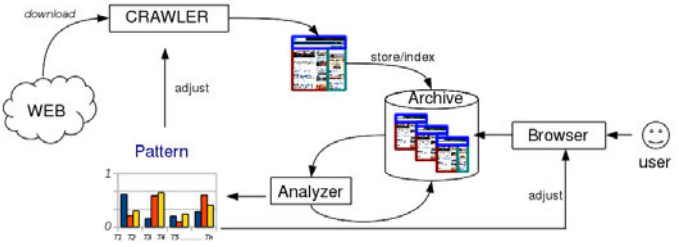


Fig. 3. Pattern-based Archiving

5 Coherence-Oriented Crawling

An ideal approach to preserve the coherence of archives is to prevent pages content from changing during the crawl of a complete collection. As this is practically impossible, we have the idea to crawl each colection during the periods of time where very little (or useless) changes are expected to occur on pages. Such periods are named *off-peak periods*. Based on discovered patterns, these periods can be predicted for each page and grouped to share a common off-peak period for the collection as shown in Figure 4. To improve the coherence, it is better to start by crawling S_2 before S_1 in order to coincide with their *off-peak* periods (Figure 4).

5.1 Crawling Strategy

Given a limited amount of resources, our challenge is to schedule collections according to their off-peak periods in a such way that it improves the coherence of the archive. We define an urgency function that computes the priority of crawling a collection S_i at time t. The urgency $U(S_i,t,\eta)$ of crawling the collection S_i at time t is

$$U(S_i, t, \eta) = [1 - \varphi(S_i, T_k, \eta)] * (t - t_{lastRefresh})$$

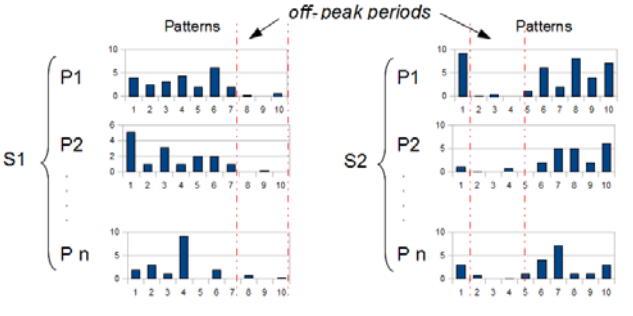


Fig. 4. Crawling collections at *off-peak periods*

- t is the current time (t $\in T_k$),
- $t_{lastRefresh}$ is the last time of refreshing the collection S_i.
- η is the number of pages considered to evaluate the coherence of A_{S_i}
- $\varphi(S_i, T_k, \eta)$ is the average of the importance of the changes predicted by patterns during the period T_k for the collection S_i.

$$\varphi(S_i, T_k, \eta) = \frac{\sum_{k=1}^{\eta} \omega_k}{\eta}$$

where ω_k is the importance of changes defined in Patt(P_i^j) ($1 \leq j \leq \eta$) at T_k.

The urgency of a collection depends on the importance of changes predicted by patterns and also on the duration between the current and the last refresh time. Less important changes occur in period T_k, higher is the priority given to crawl the collection S_i. Only the M-top collections with the highest priority are downloaded at each period T_k. The value M is fixed according to available resources (*e.g.* bandwidth, etc.). Once the M collections to be crawled are selected, the different pages are downloaded in descending order of their importance changes predicted by their patterns in period T_k. It is better to start by crawling pages with the highest changes importance because the risk of obtaining an incoherence heavily depends on the time of downloading each page. Capturing static pages at the end of crawl period does not affect the coherence of archived collection. A pseudo code of the implementation of this strategy is depicted by Algorithm 1.

Algorithm 1. Coherence-oriented Crawling

Input:
S_1, S_2,..,S_i,..., S_N - list of collections
Patt(P_i^1),Patt(P_i^2),..., Patt(P_i^j),..., Patt(P_i^n) - list of Page patterns
Begin
1. **for** each collection S_i, i=1...,N in period T_k **do**
2. compute $U(S_i, t, \eta) = [1 - \varphi(S_i, T_k, \eta)] * (t - t_{lastRefresh})$
3. collectionList.add(S_i,$U(P_i, t)$) /* in descending order of urgency */
4. **end for**
5. **for** i=1...,M **do**
6. S_i ⟵collectionList.select(i)
7. $t_{lastRefresh}$ ⟵ t
8. pageList⟵getPagesofCollection(S_i)
9. reorder(pageList, w_k) /* in descending order of changes importance */
10. **for** each page P_i^j in pageList **do**
11. download page P_i^j
12. **end for**
13. **end for**
End

6 Coherence-Oriented Navigation

In web archives, navigation, also known as surfing, is enriched with the temporal dimension. In [10], web archive navigation is represented in two different categories: horizontal navigation and vertical navigation. Horizontal navigation lets users to browse chronologically among different versions of a page, while vertical

navigation lets users to browse by following hyperlinks between pages like in the web. In this paper, we are interested in vertical navigation. Although it looks like navigation in the real web, the issues induced by web archives (temporal coherence and incompleteness) lead to broken or defected links which disable the complete navigation. As we know, it is impossible to obtain 100 % of coherence in the archive because allocated resources are usually limited and pages are too dynamic. If the system does not hold a version that was crawled exactly at the requested time, it usually returns the nearest (or recent) version. Even by finding the nearest version from the multi archives view, like in Memento framework [8], it is not sure that thuis version reflects the navigation like it was in real web. We introduce here a navigation approach that enables users to navigate through the most coherent versions.

In the remainder of the paper, the notion of collections of pages are not used anymore because while navigating in the archive, we focus on the coherence of two linked pages: (i) the source page and (ii) the destination page pointed by an hyperlink from the source page. In the following, a page is denoted by P_j and a version of the page crawled at instant t is denoted by $\mathrm{P}_j[t]$.

6.1 Informal Overview

A simple example is given in Figure 5 to better explain our coherence-oriented navigation approach. Consider a user who starts to navigate in the archive from the version of the page P_1 captured at t_q ($P_1[t_q]$). This user wants to follow the hyperlink to browse the page P_2. The closest version of P_2 before t_q is $P_2[t_1]$ and the closest version of P_2 after t_q is $P_2[t_2]$. They are the candidate destination versions. We assume that the patterns of P_1 and P_2 which describe the behavior of changes are known. These patterns are used to decide which version is the most coherent. As shown in Figure 5, the subpatterns, defined according to the periods $[t_1, t_q]$ (in red) and $[t_q, t_2]$ (in green), are extracted from patterns of P_1 and P_2. To find the most coherent version to $P_1[t_q]$, we estimate the importance of changes for each subpattern. Smaller the importance of changes predicted by subpatterns is, smaller the risk of incoherence. Thus, the group of subpatterns (a) and (b) is compared to other group of subpatterns (c) and (d) by using the importance of changes. The group of subpatterns which has the smallest total importance of changes is selected. This means that the navigation through the corresponding page versions in the selected group is more coherent. In the example, the group of subpatterns (a) and (b) has smaller importance of changes than the group of (c) and (d). Thus, the most coherent version $P_2[t_1]$ (corresponding to subpattern (b)) is returned to the user.

6.2 Formal Definitions

In this section, we give the formal definitions of our approach explained in the previous section.

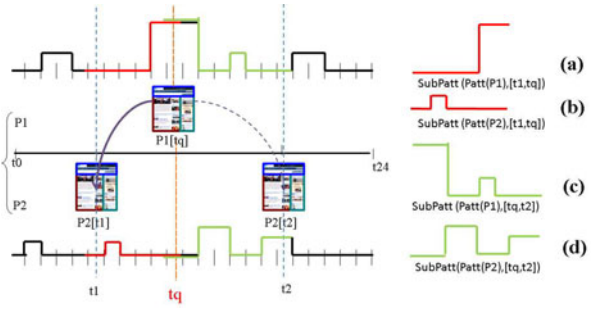


Fig. 5. Coherence-oriented Navigation

Definition 4. *SubPattern*

Given a pattern $Patt(P_j) = \{(\omega_1,T_1); ...; (\omega_k,T_k);...; (\omega_{N_T},T_{N_T})\}$, *the subpattern* $SubPatt(Patt(P_j),[t_x,t_y])$ *is a part of the pattern valid for a given period* $[t_x,t_y]$.

$$SubPatt(Patt(P_j),[t_x,t_y]) = \{(\omega_k,T_k);(\omega_{k+1},T_{k+1});...;(\omega_l,T_l)\}$$
$$where\ 1 \leq k \leq l \leq N_T \quad and\ t_x \in T_k\ and\ t_y \in T_l$$

Definition 5. *Pattern Changes Importance*

The function $\Psi(Patt(P_j))$ *estimates the total importance of changes defined in the given pattern* $Patt(P_j)$. *It is the sum of all changes importance* ω_i *of* $Patt(P_j)$.

$$\Psi(Patt(P_j)) = \sum_{i=k}^{i\leq l} \omega_i$$

Definition 6. *Navigational Incoherence*

Let $P_s[t_q]$ *be the source version where the navigation starts. Let* $P_d[t_x]$ *be the destination version* ($P_d[t_x]$) *pointed by an hyperlink. The navigational incoherence* (Υ) *between the two versions* $P_s[t_q]$ *and* $P_d[t_x]$ *is the sum of changes importance predicted by their corresponding subpatterns during the period* $[t_q,t_x]$. t_q *and* t_x *are respectively the instants of capturing the source and the destination versions.*

$$\Upsilon(P_s[t_q],P_d[t_x]) = \Psi(SubPatt(Patt(P_s),[t_q,t_x])) + \Psi(SubPatt(Patt(P_d),[t_q,t_x]))$$

where $t_q \leq t_x$

Definition 7. *Most Coherent Version*

To find out the most coherent destination version, the navigational incoherence (Υ) *between the source version and the set of the candidate destination versions are compared and the destination version with the smallest* Υ *is returned. The reason to choose the smallest* Υ *is that the probability of being incoherent depends on the importance of changes of subpatterns. In other words, if there are less changes, the source and the destination version are expected to be more coherent. The most coherent version is described as follows:*

$$MCoherent(P_s[t_q],\{P_d[t_x],P_d[t_y]\}) = \begin{cases} P_d[t_x] & if\ \Upsilon(P_s[t_q],P_d[t_x]) < \Upsilon(P_s[t_q],P_d[t_y]) \\ P_d[t_y] & otherwise \end{cases}$$

Example 1. *We take the same example of Figure 5 to explain the process. We assume that the importance of changes for the four subpatterns a, b, c, d are respectively 0.6, 0.1, 0.7, 0.6.*
The most coherent version $MCoherent(P_s[t_q], \{P_d[t_x], P_d[t_y]\})$ *is* $P_2[t_1]$ *because* $\Upsilon(P_1[t_q], P_2[t_1])$ *is smaller than* $\Upsilon(P_1[t_q], P_2[t_2])$ *where*

$$\Upsilon(P_1[t_q], P_2[t_1]) = 0.6 + 0.1 = 0.7 \text{ and } \Upsilon(P_1[t_q], P_2[t_2]) = 0.7 + 0.6 = 1.3$$

7 Experimental Evaluation

We evaluate here the effectiveness of the coherence-oriented crawling and navigation. As it is impossible to capture exactly all page changes occurred on web sites to measure the coherence, we have conducted simulations experiments based on real patterns obtained from French TV channels pages [2]. Experiments, written in Java, were conducted on PC running Linux over a 3.20 GHz Intel Pentium 4 processor with 1.0 GB of RAM. Each page is described by its real pattern and the corresponding importance of changes is generated according to this pattern. In addition, the following parameters are set: the number of pages per collection, the duration of simulation, the number of periods in patterns, the number of allocated resources (*i.e.* the maximum number of sites (or pages) that can be captured per each time period).

7.1 Coherence-Oriented Crawling Experiments

We have evaluated the coherence obtained by our *Pattern* strategy (*cf.* Algorithm 1) compared to the following related crawl policies: *Relevance* [7] which downloads first the most important sites and pages in a fixed order based on PageRank, *SHARC* [9] which repeatedly selects in a fixed order the sites to be crawled then downloads the entire site by ensuring that the most changing pages are downloaded close to the middle of the capture interval, *Coherence* [13] which repeatedly downloads sites in a circular order. Within a site, it starts by crawling the pages that have the lowest probability to cause incoherence in the archive and *Frequency* [5] which selects sites in circular order and crawls pages according to their frequency of changes estimated by a Poisson model [6].

All experiments that have been conducted to evaluate those strategies are done under the same conditions (*i.e.* a maximum of M sites can be captured at each period T).

Figure 6 shows the weighted coherence (*cf.* Section 3) obtained by the different strategies with respect to the percentage of sites crawled per period M=[10%-50%]. We varied the number η of top-pages considered to evaluate the coherence of the site (η=50%,100%). As we can see, our *Pattern* strategy, which crawls collections according to their off-peak periods, outperforms its competitors *SHARC*, *Coherence*, *Relevance* and *Frequency*. It improves the coherence by around 10 % independently of the percentage of sites crawled per period. This improvement can be observed even better if the patterns of collections are significantly different from one another.

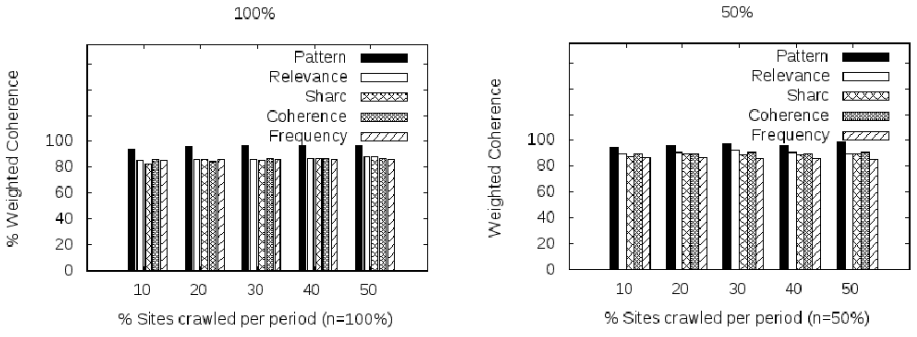


Fig. 6. Weighted Coherence

7.2 Coherence-Oriented Navigation Experiments

Similarly to the crawling experiments, we implemented our navigation approach (*cf.* Section 6) over a simulated archive based on the real patterns obtained from France TV channels pages. The experiment consists in simulating the navigation from a page source P_s to different destination pages P_d by following all outgoing links from P_s. In addition, we implemented two related navigation strategies: *Nearest* and *Recent*. The *Nearest* policy enables to navigate through the closest versions to the query time t_q. The *Recent* policy enables to navigate through the closest versions before the query time t_q. The coherence of our navigation policy *Pattern* is compared to *Nearest* and *Recent* strategies based on the definition 1 of Section 3. As we use a simulator, we know which version of the destination page is the most coherent at the beginning of the experiments. For each strategy, we count how many times the most coherent version (*i.e.* the version satisfying the coherence constraint (1)) is chosen and then this number is divided by the total number of outgoing links in the source page.

Figure 7 shows the percentage of coherent versions obtained by different strategies (*Pattern*, *Nearest*, *Recent*) with respect to the total number of outgoing links of the page source. As presented in the horizontal axis, the number of outgoing links from the page source P_s is varying from 10 to 100. We have included in brackets the percentage of cases where the nearest destination page version $P_d[t]$ to the query time t_q is incoherent with the page source version $P_s[t]$. It is important to point out that the percentage of incoherence cases presented in brackets is computed as an average obtained through several executions of simulated experiments. For example, the page source with 70 outgoing links at time t_q has about 20,7% of links where the nearest version of the destination pages are incoherent. As seen in Figure 7, our navigation policy based on patterns outperforms its competitors *Nearest* and *Recent*. It improves the coherence by around 10 % compared to *Nearest* and by around 40 % compared to *Recent*. These results are not only significant but also important since the navigation is one of the main tools used by archive users such as historians, journalists etc.

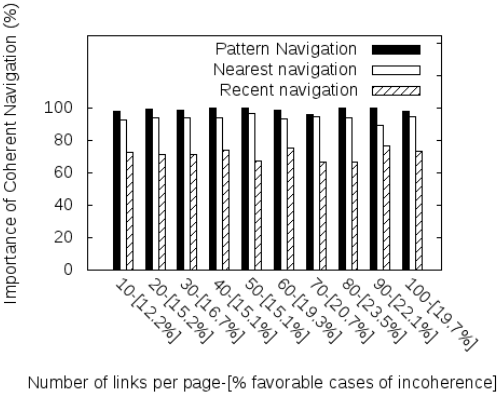


Fig. 7. Coherence-oriented Navigation

8 Conclusion and Future Work

This paper addresses an important issue of improving coherence of archives under limited resources. We proposed two solutions : *a priori* and *a posteriori*. The *a priori* solution adjusts crawlers strategy to improve archive coherence by using patterns. We have demonstrated that reordering collections of web pages to crawl according to their off-peak periods can improve archives coherence by around 10 % compared to current policies in use. Moreover, as an *a posteriori* solution, we proposed a novel browsing approach using patterns that enables users to navigate through the most coherent pages versions. Results of experiments have shown that our approach can improve the coherence during the navigation by around 10 % compared to related policies *Nearest* and *Recent*. To the best of our knowledge, this work is the first to exploit patterns to improve coherence of crawling and navigation. As a future direction, we intend to test the two proposed solutions over real data. Our challenge is to enable users to navigate through the most coherent versions at a reasonable time. Further study needs to be done to evaluate how far users can perceive coherence improvements when they navigate in the archive.

References

1. Ben Saad, M., Gançarski, S.: Using visual pages analysis for optimizing web archiving. In: EDBT/ICDT PhD Workshops, Lausanne, Switzerland (2010)
2. Ben Saad, M., Gançarski, S.: Archiving the Web using Page Changes Pattern: A Case Study. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2011), Ottawa, Canada (2011)
3. Brewington, B., Cybenko, G.: How dynamic is the web? In: WWW 2000: Proceedings of the 9th International Conference on World Wide Web, pp. 257–276 (2000)

4. Brokes, A., Coufal, L., Flashkova, Z., Masanès, J., Oomen, J., Pop, R., Risse, T., Smulders, H.: Requirement analysis report living web archive. Technical Report FP7-ICT-2007-1 (2008)
5. Cho, J., Garcia-Molina, H.: Effective page refresh policies for web crawlers. ACM Trans. Database Syst. 28(4), 390–426 (2003)
6. Cho, J., Garcia-Molina, H.: Estimating frequency of change. ACM Trans. Interet Technol. 3(3), 256–290 (2003)
7. Cho, J., Garcia-molina, H., Page, L.: Efficient crawling through url ordering. In: Computer Networks and ISDN Systems, pp. 161–172 (1998)
8. de Sompel, H.V., Nelson, M.L., Sanderson, R., Balakireva, L., Ainsworth, S., Shankar, H.: Memento: Time travel for the web. CoRR, abs/0911.1112 (2009)
9. Denev, D., Mazeika, A., Spaniol, M., Weikum, G.: SHARC: framework for quality-conscious web archiving. Proc. VLDB Endow. 2(1), 586–597 (2009)
10. Jatowt, A., Kawai, Y., Nakamura, S., Kidawara, Y., Tanaka, K.: A browser for browsing the past web. In: Proceedings of the 15th International Conference on World Wide Web, WWW 2006, New York, NY, USA, pp. 877–878 (2006)
11. Olston, C., Pandey, S.: Recrawl scheduling based on information longevity. In: Proceeding of the 17th International Conference on World Wide Web, WWW 2008, New York, NY, USA, pp. 437–446 (2008)
12. Pehlivan, Z., Ben Saad, M., Gançarski, S.: Vi-diff: Understanding web pages changes. In: Bringas, P.G., Hameurlain, A., Quirchmayr, G. (eds.) DEXA 2010. LNCS, vol. 6261, pp. 1–15. Springer, Heidelberg (2010)
13. Spaniol, M., Denev, D., Mazeika, A., Weikum, G., Senellart, P.: Data quality in web archiving. In: WICOW 2009: Proceedings of the 3rd Workshop on Information Credibility on the Web, New York, NY, USA, pp. 19–26 (2009)
14. Spaniol, M., Mazeika, A., Denev, D., Weikum, G.: "catch me if you can": Visual analysis of coherence defects in web archiving. In: 9th International Web Archiving Workshop (IWAW 2009), Corfu, Greece, pp. 27–37 (2009)
15. Teevan, J., Dumais, S.T., Liebling, D.J., Hughes, R.L.: Changing how people view changes on the web. In: UIST 2009: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 237–246 (2009)

The YUMA Media Annotation Framework

Rainer Simon[1], Joachim Jung[1], and Bernhard Haslhofer[2]

[1] AIT - Austrian Institute of Technology, Donau-City-Str. 1,
1220 Vienna, Austria
{rainer.simon,joachim.jung}@ait.ac.at
[2] Cornell University,
Ithaca, NY, USA
bernhard.haslhofer@cornell.edu

Abstract. Annotations are a fundamental scholarly practice common across disciplines. They enable scholars to organize, share and exchange knowledge, and collaborate in the interpretation of source material. In this paper, we introduce the YUMA Media Annotation Framework, an ongoing open source effort to provide integrated collaborative annotation functionality for digital library portals and online multimedia collections. YUMA supports image, map, audio and video annotation and follows the OAC annotation model in order to provide data interoperability. A unique feature of YUMA is *semantic enrichment*, a mechanism that allows users to effortlessly augment annotations with links to contextually relevant resources on the Linked Data Web.

Keywords: Annotation, Linked Data, Tagging.

1 Introduction

Annotations are a fundamental scholarly practice common across disciplines [5]. They enable scholars to organize, share and exchange knowledge, and collaborate in the analysis of source material. At the same time, annotations offer additional context. They provide explanations which may help others in the understanding of a particular item [2], or point to related material which may be useful for its interpretation. As institutions are making increasing efforts to digitize their holdings and make them available to the public over the Web [6], the role of annotations is evolving: cultural institutions are discovering the added value of user-contributed knowledge [4]. To users, adding comments, notes or tags to collection items is a convenient way to organize and personalize the information they find; or to share it with others online. To institutions, annotations can serve as a source of additional metadata which can improve search and retrieval, and helps users to discover content they wouldn't have found otherwise.

However, until now no single annotation application that can manage more than one specific media type has been widely adopted. Moreover, if annotation functionality is provided at all, it is usually based on an in-house solution, employing proprietary data models which are not interoperable with those of other

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 434–437, 2011.

systems [2]. As a result, annotation data is locked in closed silos, and usable only within the confines of a single system.

In this paper, we present the *YUMA Universal Media Annotator*[1], an ongoing effort to create an open source annotation framework for different types of multimedia content. By exposing annotations according to the principles of *Linked Data* [1], they are pulled out of institutional silos and become interoperable with other systems on the Web. YUMA is being developed in the scope of the Europeana*Connect* project[2] and is currently being showcased as part of the *Europeana ThoughtLab*[3], an online demonstration area for various initiatives carried out by partners of the *Europeana*[4] cultural heritage portal.

2 System Architecture

YUMA is based on a distributed architecture and consists of two core elements: (i) the *Annotation Suite*, a set of browser-based end-user applications for annotating content of specific media types; and (ii) the *Annotation Server*, a common "backend" service used by all of those applications. One of YUMA's key design principles is that it is designed for integration into a host environment - e.g. an online library portal - rather than to function as a standalone application. Consequently, it lacks typical portal features such as user management, and instead foresees appropriate APIs and authentication mechanisms which allow the host environment to use YUMA as an external, loosely-coupled service.

3 Annotation Suite

With the YUMA Annotation Suite the user creates "Post-It"-style annotations on digital media items. Annotations can pertain to the item as a whole, or only a part of it. At present, the suite includes tools for **images** and **digitised maps**. (The map tool is similar to the image tool, but features a "Google-Maps-like" interface for high-resolution content, and adds geographical features such as georeferencing and map-overlay.) In addition, there are prototype implementations for **audio** and **video** annotation.

The user interface offers similar functionality across all of the media types. As an example, a screenshot of the map annotation tool is shown in Fig. 1: a floating window lists existing annotations, and provides GUI elements for creating, editing, and deleting. To facilitate communication and collaboration, it is possible to reply to annotations, and to keep track of discussions around a particular media item or annotation via RSS feeds. For annotating specific areas, the tool

[1] http://github.com/yuma-annotation
[2] http://europeanaconnect.eu
[3] http://europeana.eu/portal/thoughtlab.html, also see http://dme.ait.ac.at/annotation for direct access to the annotation demonstrator
[4] http://www.europeana.eu/

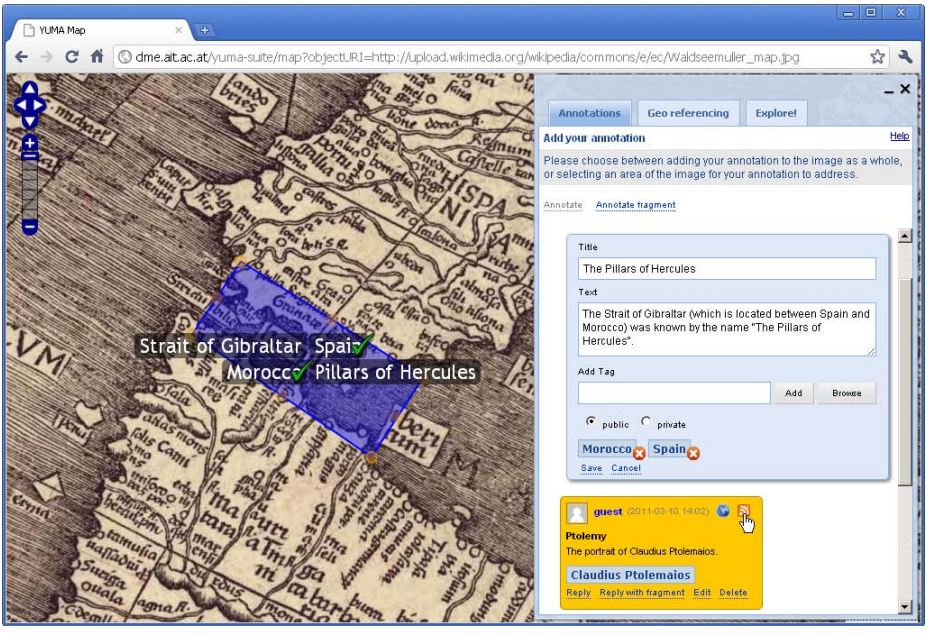


Fig. 1. YUMA Map Annotation Tool Screenshot

provides the option to mark a location or draw a polygon shape on the map. (The image tool provides the same functionality. The audio tool allows selection of a specific time range; the video tool supports both time range selection and shape drawing.)

In addition to free-text annotation, YUMA supports the notion of *semantic enrichment* of annotations by means of tags that represent resources on the Web. In contrast to free-form tags, which can be any arbitrary keyword, semantic tags are chosen from a controlled vocabulary. Besides enforcing a more coherent tagging structure, this has two added benefits. Firstly, a link to a semantic resource is not ambiguous - which is particularly valuable in the context of search and retrieval. Secondly, semantic resources can contain (or link to) more relevant information - such as descriptive text abstracts, synonymous name variants, names in different languages, or, in the case of tags referring to places, geographic coordinates. This information may not only be of interest to the user, it can also be exploited to complement traditional metadata and facilitate advanced search functionalities such as multilingual, synonym, or geographical search [3].

While it is possible to configure YUMA to work with a dedicated institutional vocabulary, YUMA's primary approach to semantic tagging is based on Linked Data: using Named Entity Recognition, the system will attempt to identify mentions of e.g. place or person names in the annotation, and suggest appropriate tags that represent resources in a Linked Data set[5]. The map annotation tool

[5] The current prototype relies on *DBpedia Spotlight* (`http://dbpedia.org/spotlight`) for this step.

also suggests tags for geographic entities inside the annotated area [4]. These suggestions are presented in the form of a tag cloud. The user can accept a suggestion by clicking on the tag, as shown in Fig. 1.

4 Annotation Server

The Annotation Server is the storage and administration backend of the YUMA Annotation Framework. It can be deployed with different relational database systems (such as MySQL or PostgreSQL). The different applications in the Suite access, store, update, and delete annotations through a REST API. The Server also offers search (through a GUI as well as through an API) and basic administration features, and provides the infrastructure for the RSS feed syndication.

Furthermore, the Annotation Server exposes annotations to the outside world as Linked Data. Each annotation is assigned a unique URI, which returns an RDF representation when resolved. To provide data interoperability, the tool relies on the OAC[6] model. OAC is an emerging ontology for describing scholarly annotations of Web-accessible information resources; and YUMA is among the first annotation solutions to implement it.

This paper presented work done for the EU-funded best practice network Europeana*Connect*, within the eContent*plus* Programme, and was also supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
2. Haslhofer, B., Jochum, W., King, R., Sadilek, C., Schellner, K.: The LEMO Annotation Framework: Weaving Multimedia Annotations with the Web. International Journal on Digital Libraries 10(1), 15–32 (2009)
3. Haslhofer, B., Momeni, E., Gay, M., Simon, R.: Augmenting Europeana Content with Linked Data Resources. In: Proceedings of the 6th International Conference on Semantic Systems, ACM, Graz (2010)
4. Simon, R., Sadilek, C., Korb, J., Baldauf, M., Haslhofer, B.: Tag Clouds and Old Maps: Annotations as Linked Spatiotemporal Data in the Cultural Heritage Domain. In: Proceedings of the Linked Spatiotemporal Data Workshop (LSTD 2010), Zurich, Switzerland, pp. 12–23 (2010)
5. Unsworth, J.: Scholarly Prumitives: what methods do humanities researchers have in common, and how might our tools reflect this? In: Humanities Computing: Formal Methods, Experimental Practice. King's College, London (2000), http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html
6. Van Der Sluijs, K., Houben, G.-J.: Metadata-based Access to Cultural Heritage Collections: the RHCe Use Case. In: PATCH 2008 - Proceedings of the 2nd International Workshop on Personalized Access to Cultural Heritage, Hannover, Germany, pp. 15–25 (2008)

[6] http://www.openannotation.org/spec

The Reading Desk: Supporting Lightweight Note-Taking in Digital Documents

Jennifer Pearson, George Buchanan, and Harold Thimbleby

FIT Lab, Swansea University
{j.pearson,g.r.buchanan,h.w.thimbleby}@swan.ac.uk

Abstract. When reading on paper, readers often write notes, fold corners or insert bookmarks without apparent conscious effort. Research into digital reading has discovered that electronic tools are far less intuitive, require significantly more attention, and are much less used. This paper introduces "The Digital Reading Desk" – a document reading interface that enhances existing digital reading interactions by adopting effective elements of paper interaction, and combining those with digital enhancements.

Keywords: Annotation, Placeholding, Digital Documents.

1 Introduction

In the physical world, the act of note-taking requires very little conscious effort. The lightweight [2] properties of paper coupled with years of learned behaviour facilitate easy manipulation and use of paper based mark-up tools. The equivalent interactions on digital document readers however, do not offer the same affordances and consequently suffer from poor rates of use when compared to their physical paper counterparts [2,3]. One contributory factor may be the over-specialised nature of digital tools, compared to print. This can be observed when using Post-Its on a physical document. Post-its can have multiple uses depending on their placement: they can sit entirely within a page (to make notes about it), protrude from the side of the document (acting as a bookmark) or even on the desk next to the document (for a note about the whole document). See Fig 1.

We are endeavouring to answer Marshall and Bly's [2] challenge to produce digital tools that mirror the apparently fluid and effortless actions seen by users of paper documents. The "Digital Reading Desk" mirrors some features of working with paper that have been absent from previous designs, and that may be essential for a truly usable system. The advantages of the design have been established by an initial user study.

2 Physical versus Digital

In terms of interaction, the way in which users can manipulate digital notes is drastically different from paper. As discussed above, there are several ways in which paper can be exploited that lack digital solutions. For example, the

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 438–441, 2011.

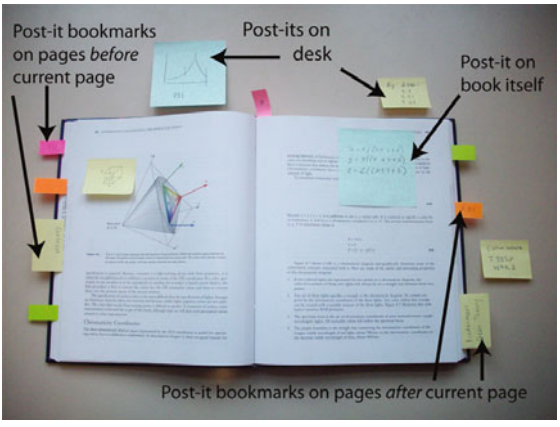


Fig. 1. An example of Post-its being used in a physical book

majority of digital reading systems do not provide a workspace comparable to a physical desk to make notes. Previous studies [1,4] have proved that the margins and indeed the space surrounding the document (the desk for example) perform an integral role in the physical mark-up process, yet thus far, little attempt has been made to integrate this feature into digital reader designs.

As discussed in the previous section, paper also easily facilitates multi- functionality within tools (e.g. Post-its that can act as notes as well as rudimentary bookmarks), whereas digital systems typically separate these functions.

3 Interaction Design

The 'Digital Reading Desk' is a system designed to overcome the problems associated with electronic reading. Our interaction is informed by the effective interactions of reading on paper. We present the PDF document as a double page spread (see Figure 2), that appears like a book on a desk – cueing the user to the potential for paper-like interaction. A key feature of the design is the incorporation of the 'virtual desk'; an area that provides an additional workspace, that can, like a physical desk, hold notes about the document. To mimic the way physical notes perform multiple functions, we have combined the note-taking and bookmarking into a single lightweight tool that will in turn reduce the time and effort required for learning. In both cases, we have therefore made an interaction that is as complete a reproduction of the book metaphor as is beneficial.

To the right of the virtual desk sit three 'piles' of inexhaustible Post-its that can change colour using the palette above. To create a Post-it, the user drags from a Post-it pile directly onto the document (see Figure 2). This approach eliminates the need for a menu system, as they are removed in same fashion, i.e. by dragging back onto the pile. The Post-its can contain text, making them mimic their physical equivalents. They can be moved, resized (to mimic folding and cutting) and 'lifted up' (to reveal the text underneath).

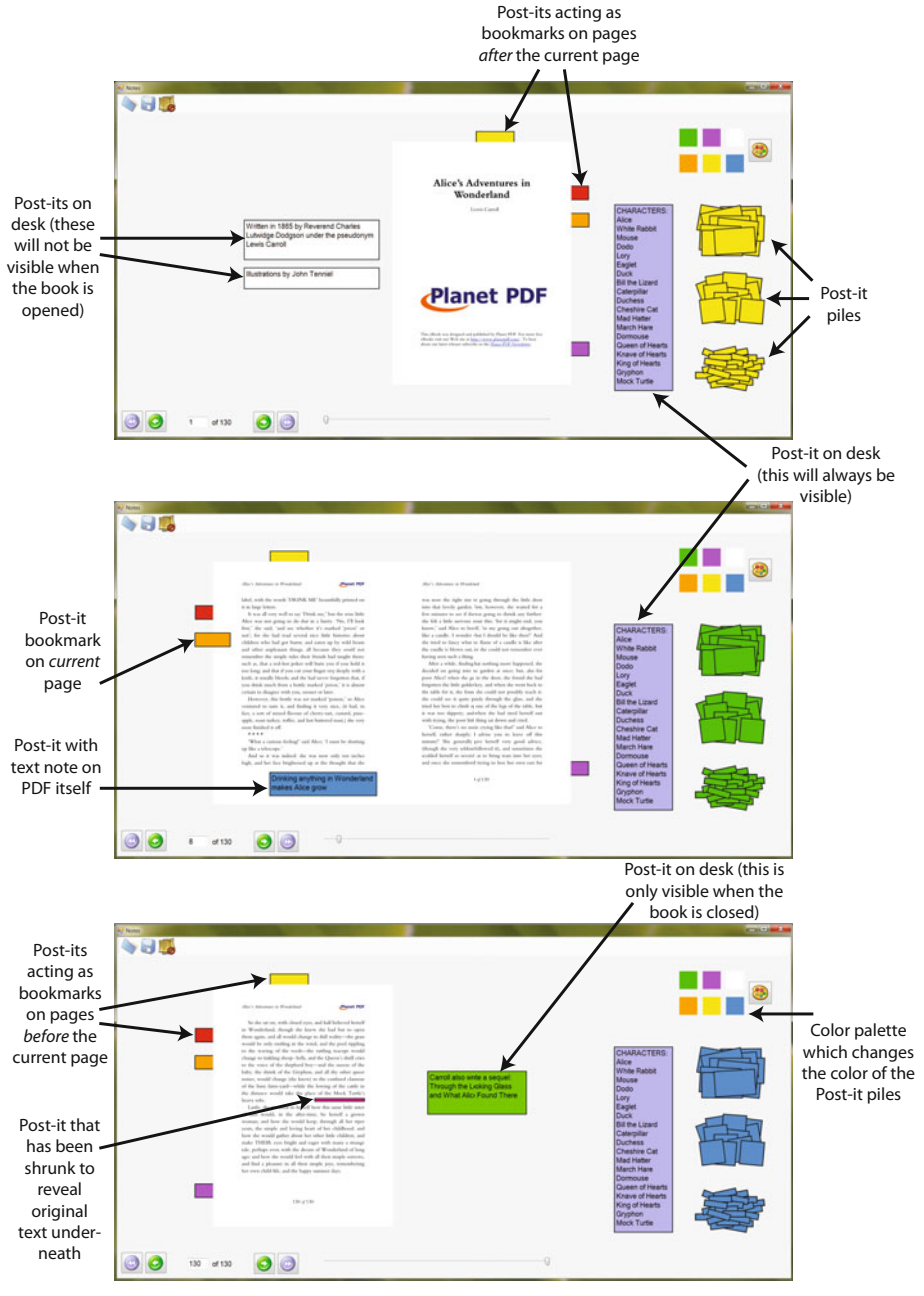


Fig. 2. Screen Shots from the Drag-and-Drop Notes System

Finally, and most importantly, to ensure the Post-its behave in the same way as on paper we wanted to ensure there were no constraints as to *where* they could be positioned. Specifically then, Post-its can be placed either:

1. Completely on the document;
2. On the desk next to, or behind (if the book is closed) the document;
3. Protruding the document which will cause it to act as a bookmark.

Thus, one tool now performs three separate functions: first, to make notes on specific pages; second, to make notes about the book as a whole and thirdly support placeholding. Post-its that also act as bookmarks not only navigate to the correct page when clicked, but also 'flip' from one side of the book to the other depending upon which page is open i.e., Post-its that bookmark pages that are sequentially *before* the current page appear on the left, and those on pages that follow *after* the current page are on the right.

4 User Study

We have conducted a sixteen participant user study on the Reading Desk, compared to two benchmark systems that replicate current "state-of-the-art" interfaces for reader software. This three-way comparison resulted in establishing clear benefits to two key features of the Reading Desk's design. First, participants subjectively approved of the provision of the 'virtual desk' area; second, the unified Post-it style tool that provides both bookmarking and note-taking simultaneously. Furthermore, participants used annotation much more extensively in open tasks with the Reading Desk than with the traditional designs.

5 Conclusions

We wish to create an effective digital system for attentive reading, that encourages users to make extensive use of annotation, as they do on paper. Our underlying hypothesis is that an extended workspace and multi-purpose tools will minimize the user's interaction effort during close attentive reading, and that the high demands of digital annotation tools contributes to their low use.

References

1. Buchanan, G., Pearson, J.: Improving Placeholders in Digital Documents. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 1–12. Springer, Heidelberg (2008)
2. Marshall, C.C., Bly, S.: Turning the page on navigation. In: JCDL 2005: Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 225–234. ACM, New York (2005)
3. O'Hara, K., Sellen, A.: A comparison of reading paper and on-line documents. In: CHI 1997: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 335–342. ACM, New York (1997)
4. Pearson, J., Buchanan, G., Thimbleby, H.W.: Improving Annotations in Digital Documents. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 429–432. Springer, Heidelberg (2009)

Metadata Visualization in Digital Libraries

Zuzana Nevěřilová

Faculty of Informatics, Masaryk University
Botanická 68a
602 00 Brno, Czech Republic
xpopelk@fi.muni.cz

Abstract. Readers in digital libraries (DL) usually do not lack information, on the contrary while browsing a DL they often struggle with too many documents. Searching and displaying search results appropriately becomes important.

This demonstration shows an experimental interface that displays search results in two forms: textual (which the readers are used to) and visual. Displaying search results as networks of similar documents, articles of the same author or articles with the same keywords often reveal new information.

Presented application is a web page with a Java Applet communicating with the rest of the page and integrated in Czech Mathematics DL website.

1 Introduction

Currently digital libraries (DL) are storages of large amounts of data, where most users (and providers) focus on searching. While browsing search results, visualization can be very helpful since the search capacity of human visual system is very large [11].

The Czech Digital Mathematics Library (DML-CZ) [1] is a large DL where the visual interface is implemented (in experimental mode) with the Visual Browser [8].

Visual Browser [8] is a general tool for dynamic (animated) visualization of RDF[1] triples. It can be used for different kinds of data thanks to its two layer architecture: the data and the *perspective of view* – a short XML description of appearance of different (classes of) nodes, edges (e.g. display articles' titles as green rectangles with round corner, display shorter edges between most similar articles). In the case of DML-CZ search results visualization, Visual Browser had not to be adapted, only a new perspective of view was created. The emphasis of the work was the conversion to RDF, suitable for visualization.

2 Metadata Visualization

Searching in the DL is in fact searching in the metadata. Metadata structure, quantity and correctness are relevant factors for successful searching in DLs. Search results are displayed in textual form, but simultaneously the same data can be browsed in visual form (see Figure 1).

From the point of view of visualization metadata are of different kinds:

[1] Resource Description Framework.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 442–445, 2011.

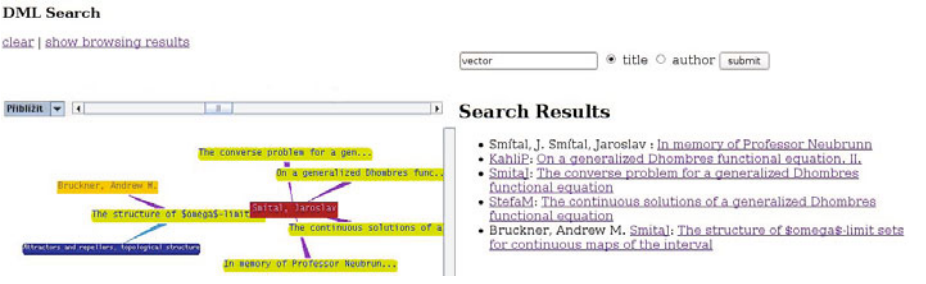


Fig. 1. Search results are displayed in both textual and visual form

- short texts (authors' names, short titles, MSC[2] categories, keywords)
- links (bibliography, similarity, classification, shared keywords)
- longer texts (full paper name, abstract)

The metadata constitute a huge, but relatively sparse network of units of different classes. In this network, nodes are labeled by short texts. Different classes of units are represented by different colors and shapes. Mapping from logical entities to their visual attributes is fully configurable in Visual Browser (via perspective of view).

Edges represent *structural* (e.g. articles in issues), *semantic* (e.g. classification of articles) or *mixed* types of relations (authors of articles).

3 Related Work

Since RDF is widely used for encoding semantic data, there exist a large number of visualization software. These applications were developed for different purposes, therefore have different features: some of them are editors as well, some of them support reasoning. A short overview is provided in this section.

Unless ABox Visualizer [7] or PGV [2] that can visualize chunks of nodes we concentrate on visualizing individual nodes and edges. However, there are still many ways to do that.

A good example of individual nodes visualization is a Web-based browser and editor SWOOP with Graph Visualization plug-in [5]. It is intended for browsing and editing OWL[3] ontologies. GrOWL [6] is another visualization and editing tool based on the underlying description logic semantics of OWL. It offers animated force directed graph layout. For displaying different classes of data in different ways, GrOWL uses filtering.

RDF Gravity [3] is another software, designed for visualizing RDF graphs. It supports filtering as well, but also allows data providers to configure the appearance of nodes and edges through configurable renderers. Node-centric RDF Graph Visualization [10] is yet another tool that displays nodes and edges in a way more suitable for printing.

[2] Mathematics Subject Classification [4].
[3] Web Ontology Language.

4 Application Background

The application follows the client-server architecture. The server stores and provides metadata, the textual and the visual interfaces are clients. Textual interface is a dynamic web page and the visual interface is a Java application running as Java applet. Moreover, these two clients are able to intercommunicate on client side via JavaScript. The whole architecture is represented by Figure 2.

DL metadata are converted from the native metadata format to RDF triples.The native metadata are stored in multiple XML files. Each serial is in a directory tree. In each directory there are metadata for the appropriate level, e.g. for serial, year, volume. The metadata have to be "escaped" (conversion of special characters) and transformed to form RDF triples such as `<articleID> dml:hasAuthor <authorID>`. Some of the native metadata were omitted (e.g. processing progress, date of scanning).

SPARQL[4] [9] is the communication language between server and clients. Each click is converted to a SPARQL query and a relatively small subgraph (SPARQL construct) is returned to the client. In case of the textual interface, a XSL transformation is used to display the data in an appropriate form. For Visual Browser RDF is its native format, therefore no more processing is needed.

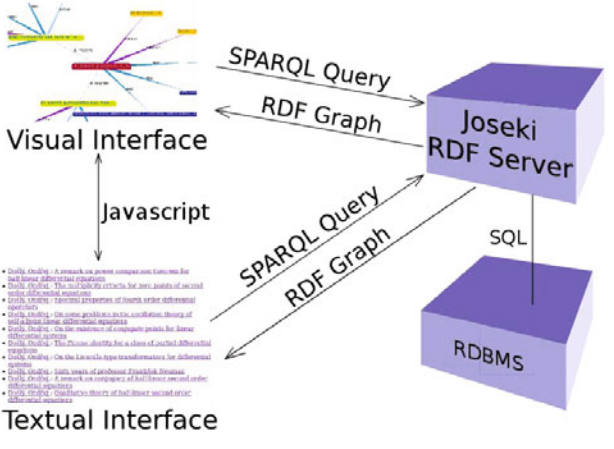


Fig. 2. Architecture of the application

5 Conclusion and Future Work

An alternative interface to a digital library is presented. The visual interface can help users to orient themselves in complex and large data. It is intended to serve together with the textual interface since users are used to textual interfaces. Currently, this interface is used (in experimental mode) to browse through search results in DML-CZ[5].

[4] Recursive shorthand for SPARQL Protocol and RDF Query Language.

[5] http://search.dml.cz

The client-server architecture was used since it is able to process large data (only a small part of the data is displayed and listed at a time). However, the quality of the visualization depends significantly on the metadata.

Future work comprises enhancements in user comfort. We plan to allow users to set their own ways to display the metadata and possibly to display different kinds of metadata (e.g. annotations).

Acknowledgments. This research has been partially supported by the grant registration no. 1ET200190513 of the Academy of Sciences of the Czech Republic (DML-CZ), and by EU project # 250503 in CIP-ICT-PSP.2009.2.4 (EuDML).

References

1. Bartošek, M., Lhoták, M., Rákosník, J., Sojka, P., Šárfy, M.: DML-CZ: the objectives and the first steps. In: Borwein, J., Rocha, E.M., Rodrigues, J.F. (eds.) CMDE 2006: Communicating Mathematics in the Digital Era, pp. 69–79. A.K. Peters, MA (2008)
2. Deligiannidis, L., Kochut, K.J., Sheth, A.P.: RDF data exploration and visualization. In: Proceedings of the ACM First Workshop on CyberInfrastructure: Information Management in eScience, CIMS 2007, pp. 39–46. ACM, New York (2007)
3. Goyal, S., Westenthaler, R.: RDF gravity (2004), http://semweb.salzburgresearch.at/ (retrieved June 1, 2011) apps/rdf-gravity/ (retrieved June 1, 2011)
4. Ion, P.: Mathematics subject classification (2010), http://msc2010.org/ (retrieved May 6, 2010)
5. Kalyanpur, A., Parsia, B., Sirin, E., Grau, B.C., Hendler, J.: Swoop: A web ontology editing browser. Web Semantics: Science, Services and Agents on the World Wide Web 4(2), 144–153 (2006)
6. Krivov, S., Williams, R., Villa, F.: GrOWL: a tool for visualization and editing of OWL ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 5(2), 54–57 (2007)
7. Liebig, T., Noppens, O.: Interactive visualization of large OWL instance sets. In: Proc. of the Third Int. Semantic Web User Interaction Workshop (SWUI 2006), Athens, GA, USA (November 2006)
8. Nevěřilová, Z.: Visual browser: A tool for visualising ontologies. In: Proceedings of I-KNOW 2005, pp. 453–461. Know-Center in coop. with Graz Uni, Joanneum Research and Springer Pub. Co., Graz, Austria (2005)
9. Prud'hommeaux, E., Seaborne, A.: SPARQL query language for RDF (2008), http://www.w3.org/TR/rdf-sparql-query/ (retrieved May 10, 2010)
10. Sayers, C.: Node-centric RDF graph visualization. Technical Report HPL-2004-60, HP Laboratories Palo Alto (April 2004)
11. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann Publishers Inc., San Francisco (2004)

Archiv-Editor – Software for Personal Data: Demo-Presentation at the TPDL 2011

Christoph Plutte

Berlin-Brandenburgische Akademie der Wissenschaften
TELOTA (The Electronic Life of the Academy)
Jägerstraße 22/23, 10117 Berlin
plutte@bbaw.de

Abstract. The Archiv-Editor is a multilingual desktop program for working with a Person Data Repository. It is developed as part of the DFG-Project Person Data Repository at the Berlin-Brandenburgische Academy of Science and Humanities (BBAW). Researchers in the humanities can enter any data related to a person, from archives, books and other sources, into the Archiv-Editor offline, and store and exchange the data with colleagues via one or more Person Data Repositories. Information about a person is not entered into a formula or table, but into an open text field and then marked with a customizable markup based on the Text Encoding Initiative. As they do not require a specific structure of statements and information, the Person Data Repository and the Archiv-Editor are open to a wide variety of research projects in Humanities and offer the infrastructure to combine and integrate data from divergent fields and research perspectives.

1 Person Data Repository at the BBAW

The project "Construction of a repository for biographical data on historical persons of the 19th century" – short form: Person Data Repository – enhances the existing approaches to data-integration and electronically supported biographical research. It investigates connecting and presenting heterogeneous information on persons of the "long nineteenth century" (1789–1914). The project's aim is to provide a de-centralized software system for research institutions, universities, archives, and libraries, that allows combined access to biographical information from different data pools. The project has installed the first repository at the BBAW and a second repository in Rome for an international group of researchers, and is developing further cooperation with projects in Bonn and Bamberg.

To structure heterogeneous biographical data and to deal with non-predefined structures, the project pursues a novel approach. It defines a person as a compilation of all statements (considered as "aspects of a person") concerning that person. Thus it is possible to display complementing as well as contradicting statements in parallel, which meets one of the basic challenges of biographical research.

Further Information about the project: http://pdr.bbaw.de

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 446–448, 2011.

2 The Archiv-Editor Software

The current version of the Archiv-Editor is based on experiences made with a first version of the software since 2005 at the BBAW. The first stable version of Archiv-Editor 2.0 was released in March 2011. Further development of the software during the next two years is already financed and will very probably be prolonged.

2.1 Concept and Main Tasks

The Archiv-Editor is the central interface for researchers to enter data into the Person Data Repository database, to modify it, search and find relations between persons. It allows working offline in order to support working directly in archives without internet access. It integrates common international identifiers such as PND, LCCN or VIAF into the workflow to enhance exchangeability, and is designed for various flexible perspectives on the data related to a person, a source, a place, a time, and customizable, user-defined categories. Markup information and relation-statements allow the researcher to describe and classify relations between persons and thus to cover wide social networks in detail.

2.2 Simplifying XML

Within the Person Data Repository data is stored in XML to allow long-term archiving and greater flexibility of textual markup based on the proposals of the Text Encoding Initiative and on the MODS standards for all types of sources. The XML is fully encapsulated in order to simplify the work with the Archiv-Editor and to avoid mistakes. XML markup is set by inserting color-markups that allow the user to see the marked categories in a popup menu on mouse-over.

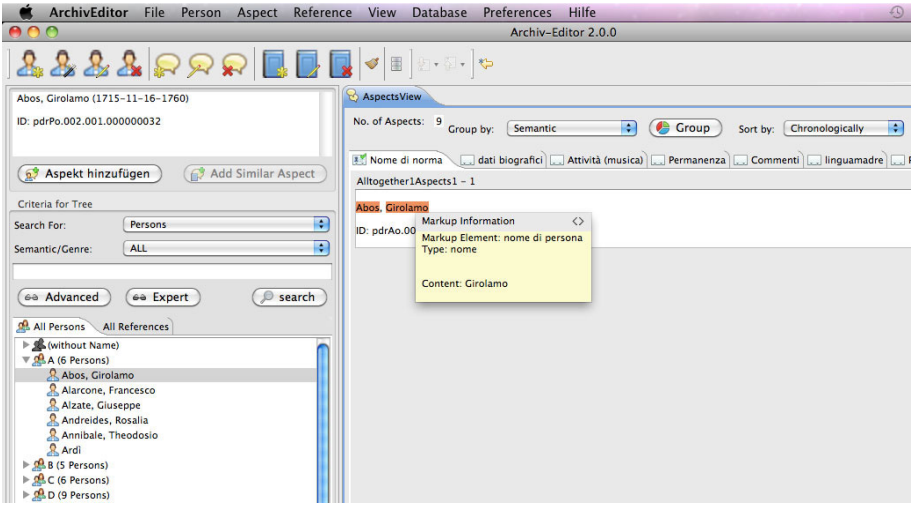
2.3 Functionality and Technology

- Persons, statements about persons and sources can be edited, searched and sorted by a wide variety of customized criteria
- Persons can be linked with international identifiers such as PND, LCCN, VIAF and others
- Sources are based on source-type (e.g. book, article) that can be edited and customized and stored according to MODS standard
- A full range of search functions such as faceted search are supported for all data objects
- Data-Cleaning tools allow the retrieval of duplicates and erroneous data
- Preferences allow the user to set styles, display-names, favorite markups and often-used functions according to her individual requirements
- Export into PDF, HTML, TXT, XML is supported for predefined as well as custom style sheets
- Import functions allow to exchange data beyond the synchronization with the central data repository
- Language Support for English, German, Italian and French is provided

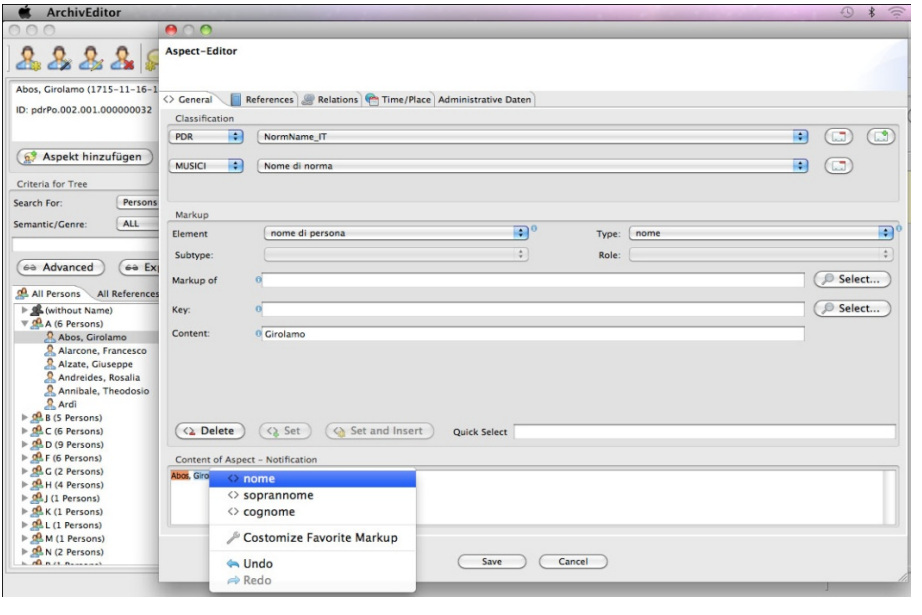
The Archiv-Editor is open source and written in Java. It is provided for Windows, Mac OS and Linux and based on the Eclipse RCP architecture which allows

extensibility and customization through plugins. This includes also a complex help system and a software update mechanism. Scenarios to integrate the Archiv-Editor into other Research Environments through plugins are currently being examined.

3 Screenshots



Main Perspective – Personal Tree and Aspects View with Markup



Editing a statement about the name of a person and setting markup

The MEKETREpository - Middle Kingdom Tomb and Artwork Descriptions on the Web

Christian Mader, Bernhard Haslhofer, and Niko Popitsch

University of Vienna, Faculty of Computer Science
{firstname}.{lastname}@univie.ac.at

Abstract. The MEKETREpository (MR) allows scholars to collect and publish artwork descriptions from Egypt's Middle Kingdom (MK) period on the Web. Collaboratively developed vocabularies can be used for the semantic classification and annotation of uploaded media. This allows all users with system access to contribute their knowledge about the published artworks. All data, including annotations and vocabularies, are published as Linked Data and can be accessed and reused by others. This paper gives an overview of MR's functionalities and the current state of our work.

1 Introduction

The MEKETRE project[1] is an interdisciplinary project conducted by the University of Vienna's Egyptology and Computer Science departments. The aim of the project is to collect and study digital representations of two-dimensional artworks (reliefs and paintings) stemming from tombs built during the MK in ancient Egypt.

The technical aim of the project is to provide a system (the MR) that enables researchers in the Egyptological domain to easily upload, classify and annotate digital versions of selected art items from that period. The collected data should be accessible in two representations, a human readable and a machine friendly version. Therefore, the data contained in the MR, except for media objects restricted by license, is available as Linked Open Data [1] on the Web.

Since its start in November 2009, the MEKETRE project has already collected comprehensive and detailed information on MK tombs and 2-dimensional art items. Much of this information is already available in the MR, and first experiences indicate acceptance by scholars from the Egyptology department.

[1] The project is funded by the Austrian Science Fund (FWF) and is scheduled for 3 years until late 2012. Further information is available at http://www.meketre.org

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 449–452, 2011.

2 Motivation

Although several databases provide information about Old Kindom (OK) art items[2], comparable sources for the MK are still missing. A problem with existing databases from the OK domain is, that for comparative studies the art items need to be described at a very high level of detail that is not present in existing databases. Further, existing databases offer limited search and retrieval functionality and most of them are closed repositories that do not provide machine friendly access to the collected raw data. The MR addresses these problems by

- Supporting description of art-items at many levels of detail.
- Opening the repository and making its content accessible on the Web in human readable and machine friendly form.
- Utilizing common standard ontologies for structuring both queries and their resulting data.

3 Methodology

For organizing artworks in the MR we distinguish between two types of first class objects: tombs and themes. Tombs are described by various properties, like the necropolis they are located at, a unique identifier, an owner (buried person), media objects (images), etc. Themes are 2-dimensional art items usually located at a tomb's walls that depict a certain detail of the life in ancient Egypt. Every theme may contain various details that need to be described and classified, e.g., because they also appear in other themes located at different tombs.

The MR allows Web users to contribute knowledge about these first-class objects using an annotation mechnism. It is possible to highlight an (rectangular- or polygon-shaped) area of interest (Fig. 1). This highlighted region can then, for example, be classified using a controlled vocabulary or described by free-text. Each first-class object can be referenced by any number of annotations from every user of the repository with write-access. Thus it is possible to annotate an arbitrary number of visual details in the depictions of a MK tomb or theme.

In the course of the project, controlled vocabularies are collaboratively developed by scholars from the egyptological domain. So far, no comprehensive standard vocabulary for describing art items exists that is widely accepted by the Egyptological community. Our approach is to support creation of such a vocabulary in a collaborative way with the goal to reflect the opinions of all contributors. This is done by integrating an online tool for vocabulary development (PoolParty[3]) into the MR user interface that supports collaborative development

[2] A database of OK scene details has been created by the Oxford Expedition to Egypt in close cooperation with the Archaeology Data Service. It is accessible online at http://ads.ahds.ac.uk/ and based on [2]. Also dealing with OK scenes is the Leiden Mastaba Project (see http://www.peeters-leuven.be/boekoverz.asp?nr=8170 and [3]) which is available on CDROM only.

[3] http://poolparty.punkt.at/

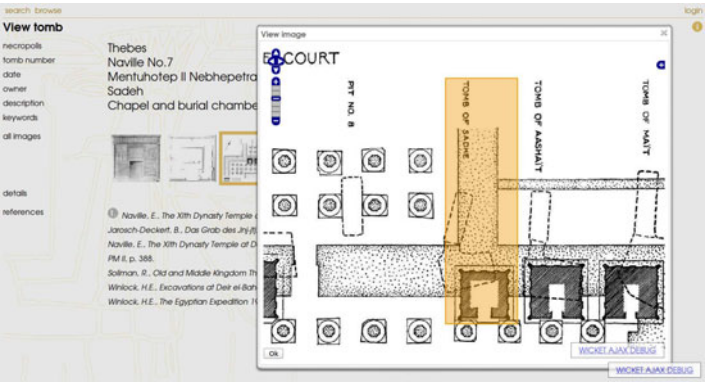


Fig. 1. An annotation (orange overlay) that shows the location of a tomb

of vocabularies that are made available on the Web, using the SKOS[4] (de-facto) standard. Besides the integration of this vocabulary development tool, we integrated the Web Reference Database (refbase[5]) in order to be able to manage, cite and reference literature that further describes a MK tomb, theme or media object.

The content of the repository will further be periodically replicated to the University of Vienna's long-term archiving solution PHAIDRA[6]. This allows for secure (versionized) storage with high availability and constant citability.

Finally, all data (art item descriptions, annotations, references, vocabularies, etc.) stored in the MR are made available as HTML and RDF (for human, respectively, machine consumption) using the Linked Data [1] principles. By this, these data can be easily integrated into other systems and data sets.

4 Current Implementation Status and Database Content

A beta version of the MR is available online[7]. Up to now, almost 120 tombs have been entered, including comprehensive metadata like date information, free-text description, ∼400 images and ∼700 annotations. Furthermore, an initial thesaurus for classifying themes has been created as well as a list of standardized terms to be used, for example, as person names and keywords. The process of entering tombs, themes and according metadata as well as creating controlled vocabularies (at the moment dozens of terms) is currently ongoing.

[4] http://www.w3.org/2004/02/skos/
[5] http://www.refbase.net
[6] https://phaidra.univie.ac.at/
[7] Visit http://www.meketre.org/repository/ using a current version of Firefox or WebKit-based browser. An RDF representation of the data can be downloaded by dereferencing the preliminary URL http://www.meketre.org/triplify/

MR supports full-text and faceted search based on Apache Lucene. We further support the OpenSearch[8] format to enable a simple integration of our search results into other applications. A preliminary draft of the MEKETRE RDF vocabulary for describing various aspects of art items (e.g., their location, depictions, annotations, temporal classification) exists. This vocabulary makes use of other well-known ontologies such as Dublin Core and FOAF. Triplify[9] is used to expose the data from the underlying relational database as Linked Data.

Long-term archiving support is in its beginning state, currently supporting the export of tombs including all their metadata and images.

5 Intended Demo and Future Work

In our demo we will present the system's user and data interfaces. Interaction with the Web frontend is shown by performing tasks like browsing, searching and adding new items. The adoption and usage of controlled vocabularies will be motivated and the vocabularies developed in the MEKETRE context will be introduced. Their practical application is illustrated by using them for content annotation. Accessing the machine friendly representation of the data constitutes another major point. The advantages of publishing content as linked data will be discussed, supported by examples of obtaining RDF data from the MR. The current status of the MR's interaction with PHAIDRA is also subject of the demo, providing details on the utilized replication approach.

Future work will encompass a detailed survey of existing ontologies and vocabularies and an analysis on their applicability in the MEKETRE context (e.g., the OAC[10] data model). Integration with tools specialized in supporting the development of SKOS vocabularies is another milestone.

We believe that due to its accessibility features (e.g., Web enabled, read access for everyone, stable URIs for artworks) and reliance on established standards for data annotation and publication, the MR will constitute an open and so far unparalleled source of knowledge about the MK, forming a basis for further research in the field.

References

1. Bizer, C.: The emerging web of linked data. IEEE Intelligent Systems 24, 87–92 (2009)
2. Harpur: Decoration In Egyptian Tombs. Studies in Egyptian Archeology. Routledge, New York (1987)
3. van Walsem, R.: Iconography of Old Kingdom Elite Tombs: Analysis and Interpretation. In: Theoretical and Methodological Aspects. Mededelingen en Verhandelingen Van Het Vooraziatisch-Egyptisch Genootschap Ex Oriente Lux. Peeters (2006)

[8] http://www.opensearch.org
[9] http://triplify.org/
[10] Open Annotation Collaboration, http://www.openannotation.org/

NotreDAM, a Multi-user, Web Based Digital Asset Management Platform

Maurizio Agelli, Maria Laura Clemente, Mauro Del Rio, Daniela Ghironi, Orlando Murru, and Fabrizio Solinas

CRS4
Building 1, Science and Technology Park Polaris
Piscina Manna, 09010 Pula (CA) - ITALY
{agelli,clem,mauro,dghironi,orlando,fabrizio.solinas}@crs4.it,
http://www.crs4.it

Abstract. In this work we present an overview of NotreDAM, an open source Digital Asset Management platform targeted to the mid-market segment. NotreDAM provides a web-based multi-user application environment for uploading, annotating, cataloguing, sharing, searching and retrieving digital resources such as videos, audios, images and documents. NotreDAM main advantages are: XMP metadata support, user-defined workspaces and catalogs, scalable processing of resources, a scripting engine extendible through plugins and a REST API for integration with third party applications. The demo will showcase the capabilities of the platform through a typical user session.

Keywords: DAM, Digital Asset Management, metadata, XMP, resource processing, open source, web based, demo.

1 Introduction

Modern econonomies strongly rely on digital assets (such as images, documents, audio and video files) for fullfilling their vital functions. The growth of creative industries is driving the need for solutions able to manage large digital collections throughout their whole life cycle. The proliferation of user-generated content is making this need even more acute. The term "Digital Asset Management" (DAM) encompasses tools and practices for cataloguing, organizing and preserving digital assets, so that they can efficiently serve the needs of all users of the value chain.

NotreDAM is an open source Digital Asset Management platform developed by CRS4 within the DistrICT Lab, a research and development initiative funded by the Sardinian Regional Operational Programme. NotreDAM is distributed under the terms of the GNU General Public License (GPL) Version 3 and can be downloaded from the project web site (http://www.notredam.org), where an online demo is available too.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 453–456, 2011.

2 NotreDAM Platform

The NotreDAM Digital Asset Management platform was designed to address the needs of the mid-market segment, such as small workgroups, stock agencies, digital libraries, niche applications, etc.. It provides a multi-user environment, accessible from most web browsers, where users can upload, annotate, catalog, share, search and retrieve resources. The following paragraphs provide an overall description of NotreDAM features.

2.1 Digital Items

The term *digital item* identifies a structured representation of a digital asset within NotreDAM. An item includes a set of resources and a set of metadata. Resources are blocks of binary information which encode the actual content of the asset. NotreDAM is able to handle videos, images, audios and documents. It supports many encoding formats, among which BMP, GIF, JPEG, TIFF, PDF, PNG, DV/AVI, Mpeg-1 Video, Mpeg-2 Video, H.264, H.263, Xvid, AAC, AIFF, Dolby AC3, Mpeg-1 Audio, OGG Vorbis, WAV.

Metadata, i.e. structured information describing the properties of digital assets, play a fundamental role in organizing and retrieving digital resources. This is particularly true in the digital media industry, due to the non-textual nature of content. NotreDAM metadata are based on XMP [1] [2] [3], an open specification which allows to embed metadata into the resource files themselves. Although created by Adobe, XMP is an open technology based on W3C's RDF and freely available to developers (Adobe provides the XMP Toolkit under a BSD license). It has become an increasingly adopted standard both among creative industries and open source communities.

NotreDAM also supports IPTC, Dublin Core, EXIF, PLUS and Creative Commons, which are encompassed by available XMP namespaces and extensions. As far as Mpeg-7 is concerned, high level attributes (part 5 Multimedia Description Schemes [11]) can be mostly mapped to XMP properties. Low-level attributes ([9] [10]) could be supported by a dedicated XMP extension. However, there is no commonly agreed mapping to RDF [12].

NotreDAM can seamlessly manage different renditions of a digital item. Renditions are distinct views of the same item, semantically equivalent but different in many respects: e.g. size, watermarking or encoding format. Each rendition is associated to a distinct resource, which can be uploaded by the user or generated internally. Each item is treated as a single object, despite the fact that it may include different renditions.

Metadata can be associated either to the item or to a single rendition. For example, a common case where it is useful to define rendition-level metadata is to attribute a Creative Commons license to a low-quality or watermarked rendition, while retaining the full rights on the original resource.

2.2 Workspaces

Workspaces are working environments where NotreDAM users can manage digital assets. The purpose of workspaces is twofold: (1) to provide a set of tools

customizable to meet the needs of specific user domains, (2) to act as collaboration spaces, where users can share items. Each workspace provides a catalog to organize items into a user defined taxonomy. This catalog-based classification is totally separated from the metadata associated to the objects, however it can be used as a shortcut for setting metadata. Although the metadata model already provides a full infrastructure for describing digital assets, the catalog is useful for creating a relation among items and the actual operating context where they will be used. Each workspace can define its own set of renditions and a metadata abstraction layer, enabling users to adopt their own descriptors, which will then be mapped to XMP properties.

Managing digital assets generally involves a significant amount of processing for carrying out operations such as harvesting metadata, transcoding resources or generating new renditions. NotreDAM provides a scripting engine for applying a sequence of actions on a batch of digital items. A default set of actions is provided, which can be extended through user-defined plugins.

3 Architecture

NotreDAM is a rich internet application, whose architecture is shown in Fig.1. Users access NotreDAM either through a browser or through third party applications that can be easily integrated via a REST API.

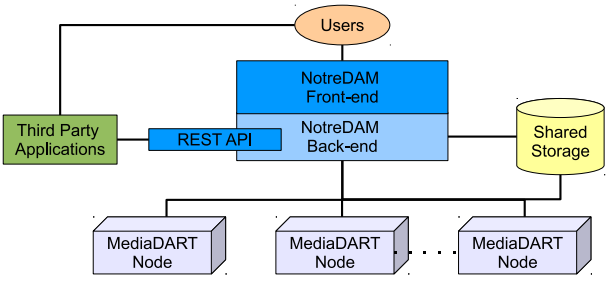


Fig. 1. NotreDAM architecture

Resource processing relies on MediaDART [4], a scalable cluster of computers communicating through AMQP [5] protocol, using the RabbitMQ [6] implementation. The front-end is based on ExtJS [8], a Javascript framework, the back-end is based on Django [7], a MVC framework for web applications. A NFS file system containing the resources is shared between the back-end and the MediaDART nodes.

4 Demo

The demo, accessible through a browser, will showcase the capabilities of the platform through a typical user session. In particular, it will allow participants

to evaluate most of the features of NotreDAM, such as uploading resources, editing metadata, using the catalog, creating a script, working with renditions. A short demonstration of the management interface can also be carried out.

5 Conclusion and Future Work

NotreDAM presents significant advantages over similar open source tools. Firstly, workspaces provide context-oriented classification environments where users can share digital assets. Secondly, in spite of the fact that a digital asset can have different renditions and metadata views or can be added to different catalogues, it is modeled to always appear as a single object which can be exported in compliance with open standards. Finally, resource processing can be split up across multiple computers, for improving performance and scalability. Future work will add new functionalities, such as a workflow engine and the capability to model real objects (e.g. artwork) in the catalog.

References

1. XMP Specification Part 1, Data Model, Serialization and Core Properties, Adobe Systems (July 2010), http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf
2. XMP Specification Part 2, Additional Properties, Serialization and Core Properties, Adobe Systems (July 2010), http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart2.pdf
3. XMP Specification Part 3, Storage in Files, Serialization and Core Properties, Adobe Systems (July 2010), http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart3.pdf
4. MediaDART framework, http://www.mediadart.org/
5. AMQP Specification, http://www.amqp.org/confluence/display/AMQP/AMQP+Specification
6. RabbitMQ platform, http://www.rabbitmq.com/
7. Django framework, http://www.djangoproject.com/
8. ExtJS framework, http://www.sencha.com/products/extjs/
9. ISO/IEC 15938-3 Information technology - Multimedia content description interface - Part 3: Visual
10. ISO/IEC 15938-4 Information technology - Multimedia content description interface - Part 4: Audio
11. ISO/IEC 15938-5 Information technology - Multimedia content description interface - Part 5: Multimedia Description Schemes
12. Multimedia Vocabularies on the Semantic Web, W3C Incubator Group Report (July 2007), http://www.w3.org/2005/Incubator/mmsem/XGR-vocabularies-20070724/

A Text Technology Infrastructure for Annotating Corpora in the eHumanities

Thierry Declerck, Ulrike Czeitschner, Karlheinz Moerth, Claudia Resch, and Gerhard Budin

Institut für Corpuslinguistik und Texttechnologie (ICLTT)
Zentrum Sprachwissenschaften, Bild- und Tondokumentation
Österreichische Akademie der Wissenschaften (ÖAW)
Sonnenfelsgasse 19/8, 1010 Wien, Austria
{thierry.declerck,karlheinz.moerth,ulrike.czeitschner,
claudia.resch}@oeaw.ac.at, gerhard.budin@univie.ac.at

Abstract. We present in this demonstration paper the actual text technology infrastructure we have been establishing for annotating with linguistic and domain-specific information – the personalized death – a corpus of baroque texts (in German) belonging to the genre "Danse Macabre". While the developed and assembled tools are already covering the automatic treatment of various lexical aspects of such texts, and are also supporting the manual annotation of the corpus with concepts related to the personalized death, we are currently extending our work with the integration of methods and tools for automating the annotation procedure. The goal of our project is to offer the philologist, historian or the interested public an improved access to this kind of corpora, allowing for example for topic based queries and navigation.

Keywords: Historical and Literary corpora, NLP, Semantic Annotation.

1 Introduction

A goal of our work is to provide high-quality linguistic and semantic annotation of textual data in the Digital Humanities & Cultural Heritage fields. For this purpose, we are handling various types of data, such as lexicons (dictionaries), annotated corpora, grammars and domain specific terminological and semantic resources. As a special case, we are dealing with a Viennese Danse Macabre corpus, built at ICLTT which consists of a digital collection of printed German texts dating from 1650 to 1750. It has been designed to allow research on a wide variety of literary, cultural and linguistic topics, with the aim to bring together traditional philological expertise with up-to-date text technology, which could be used in the context of a digital library.

In the following sections, we briefly first present the corpus and its associated resources, then the processing steps we implemented so far, before sketching some future steps.

2 The Corpus

The corpus is made up of selected digitized German texts belonging to the so-called *dance of death* and *memento mori* genres. The printed sources consist of both prose

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 457–460, 2011.

and verse text and contain many illustrations. Some texts are ascribed to the famous Viennese theologian Abraham a Sancta Clara (1644-1709).

Although personifications of violent death were very popular at that time, the texts also allow other conceptual representations of death and dying. A primary aim of the project was thus to apply semantic annotations for rendering such conceptual diversity more easily discernible and to support access to comparable digital resources.

The domain specific markup, based on an in-house developed taxonomy, which has been manually applied to the digital texts, follows a TEI conformant tag-set[1]. Instances of death as a personified entity are marked-up in the following manner:

```
<rs type="death" subtype="figure">Mors, Tod, Todt</rs>
<rs type="death" subtype="figureAlternative">General Haut und Bein, Menschenfeind</rs>
<rs type="death" subtype="attribute">knochenreich, ohnartig, vnersättlich</rs>
<rs type="death" subtype="activitySpeechAct">Auch selbst die Cron / ich nicht verschon.</rs>
<rs type="death" subtype="event">den Geist aufgeben, aus der Welt schleichen</rs>
```

The results of this basic annotation phase allow already navigating in the full text, as can be seen in the screen shot in Fig 1 that displays the actual browser based interface.

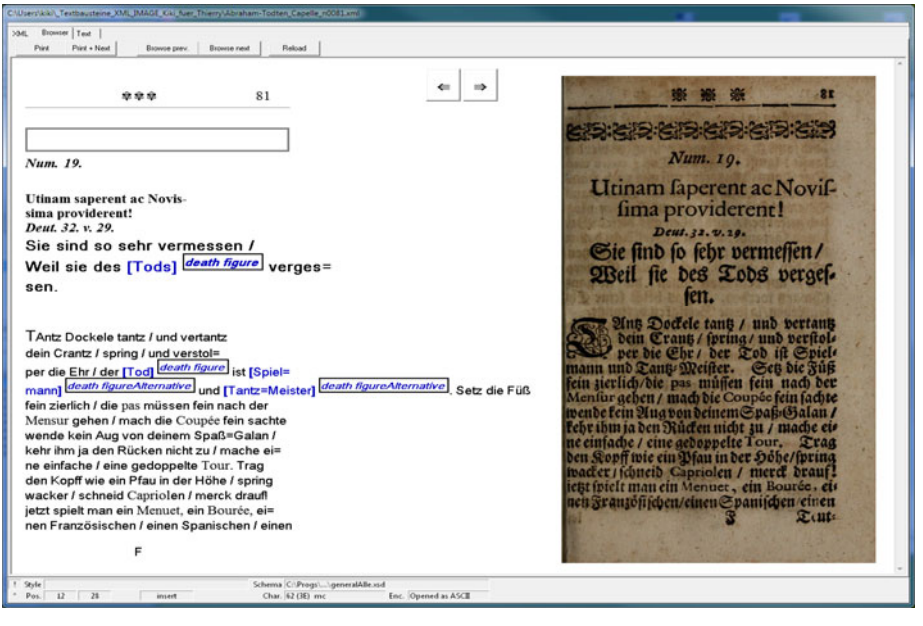


Fig. 1. The actual interface: dual display with digital text, annotations and facsimile

[1] We use TEI P5 (2008). See http://www.tei-c.org/Guidelines/P5/ for more details.

3 Lexical Processing

Lexical processing of the digitalized texts is mainly done by the ICLTT's tool "*corpedUni*", which is an editor that can process and manipulate a large number of XML documents. Additionally, one can define and protocol corpora workflows *with corpedUni*[2]. For supporting lexical processing, *corpedUni* makes use of a built-in tokeniser that generates TEI conformant *w*-elements (or "word forms"), which are also designed to hold part-of-speech (PoS) and lemma information[3].

Our corpus contains a large number of orthographic variants, due to both historical changes and "errors" of the OCR process, when applied to those specific textual sources (see again Fig 1 for an example of such a text). To identify such lexical variants, we applied a simple method, consisting in three steps: (a) create a complete list of word forms, (b) perform automatic normalizations on the character level, and (c) compare each word forms with all the others in the list making use of the so-called DICE coefficient. The result of this process is a list of word forms - including lemma and PoS information, together with all similar items in the list (see Fig. 2. below):

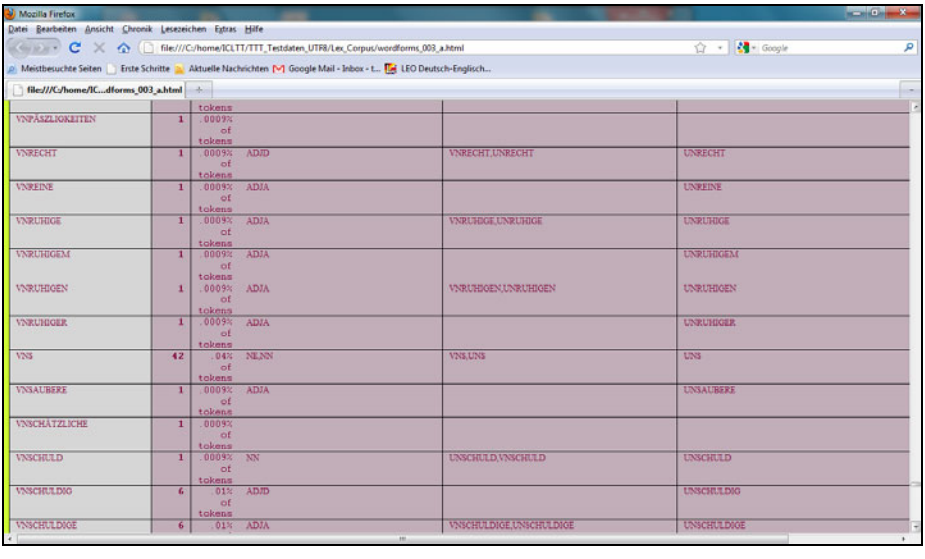


Fig. 2. Extracted word forms, with lemma and PoS information, and their variants

4 Syntactic Annotation

Syntactic processing can take place on the basis of the list of word forms, together with the lemma and PoS information, and add further linguistic annotations to the

[2] See http://www.aac.ac.at/text_tech_tools.html for more details on corpedUni.

[3] We apply TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ for adding PoS and lemma information to the word forms.

text, like phrasal information (noun phrases, etc.) and grammatical functions (like subject, predicate or object, etc.). This allows to state precisely if an item assigned to death is used in the subject position of a sentence, leading to the heuristics that death is involved as "agent" in a specific event.

We use here the NooJ tools[4], which allow the import of the manual XML annotation (see section 1) and the lexicons extracted from the corpus (see section 2), also taking into account historical variations. It is for example very easy to represent in the computational lexicon of NooJ the fact that "Clarinette" and "Klarinette" are (historical) variants, so that using a parser trained on modern German can also cope with word forms that do not correspond to the actual orthography. The example below shows how the variant "Clarinette" is put into relation with the actual word "Klarinette", sharing the same flectional (FLX) and semantic (SEM) properties.

klarinette,N+FLX=Fem_E+SEM=Music
clarinette,klarinette,N+FLX=Fem_E +SEM=Music

NooJ allows defining grammars for parsing syntactically the corpus and for adding semantic annotation manually and automatically.

5 Conclusion and Future Work

We presented a set of tools that are capable of producing various levels of annotation and indexation of digitized text selected within a certain domain and within a certain era: the Danse Macabre Corpus.

Apart from making such a digitized corpus easily available to the interested public, we also plan to establish links to documents dealing with specific historical events of this era and also to establish links to documents of this period dealing with health and disease: Our Danse Macabre Corpus contains numerous instances of names of diseases, and linking those to other medical documents of this time could help to establish a better understanding of intertextual processes.

It is planned to make the corpus accessible through a web interface, which will grant users simultaneous access to full text, facsimiles and additional material.

Our main task consists now in implementing strategies for a semi-automatic semantic annotation of the corpus, and to evaluate our approach.

References

1. Eybl, F.M.: Abraham a Sancta Clara. In: Killy Literaturlexikon. Hrsg. v. Wilhelm Kühlmann. Band 1, Seite 10-14. de Gruyter, Berlin (2008)
2. Goodwin, W., Sarah, Bronfen (Hrsg.), E.: Death and Representation. Baltimore. The Johns Hopkins University Press, London (1993)

[4] http://www.nooj4nlp.net/pages/nooj.html

An Application to Support Reclassification of Large Libraries

Kai Eckert and Magnus Pfeffer

University Library
University of Mannheim, Germany
{lastname}@bib.uni-mannheim.de

Abstract. In this paper, we describe a software application that was developed and is successfully applied at the Mannheim University Library to manually reclassify about 1 million books in a very efficient manner by supporting various different working strategys and by using information from several sources.

1 Background

A common example for the reclassification of large library collections is a switch to the Library of Congress Classification scheme. Using the classification numbers and cutter information provided by OCLC significantly reduces processing time for newly aquired titles and is a logical next step for libraries that use OCLC services for their catalogues. The situation in Germany, however, is different. Historically, no single institution with an authority comparable to the Library of Congress or the British Library emerged. As a consequence, libraries formed union catalogues to facilitate collaboration and reuse, but classification systems only slowly gained acceptance, with many libraries using self-designed or adapted classification systems.

In recent years there has been an increase in the construction of new library buildings that are designed to house the merged contents of several smaller libraries. Having multiple call number systems in an integrated library building severely obstructs the use of materials, and a switch to a unified classification system is deemed necessary. The decision on which classification system to choose is often based on the proportion of already classified titles in the union catalogs. Many libraries choose the Regensburg union classification (abbreviated as RVK) which in recent years has gained some prevalence in Germany [4], so that classification information contributed by other libraries can be reused. But still, RVK class numbers are available for only about 50% of the titles in German union catalogs. This is an additional handicap on top of the already challenging task of reclassifying large libraries as described by recent reports [1,3].

In this demo we introduce an application that is intended to assist library staff with the reclassification task, i.e. the application of class numbers to individual titles. It is actively deveolped at the library of the university of Mannheim, Germany and is currently applied by several librarians who are working on the ongoing reclassification of the roughly 1 million books available in open access areas[1].

[1] The books are located in four libraries in different buildings on the campus. The closed stacks of the library hold another 1 million books.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 461–464, 2011.

2 The Application

The main paradigm of the classification application is to support the users in all possible ways while still keeping the whole process strictly intellectual and under the full control of the users.

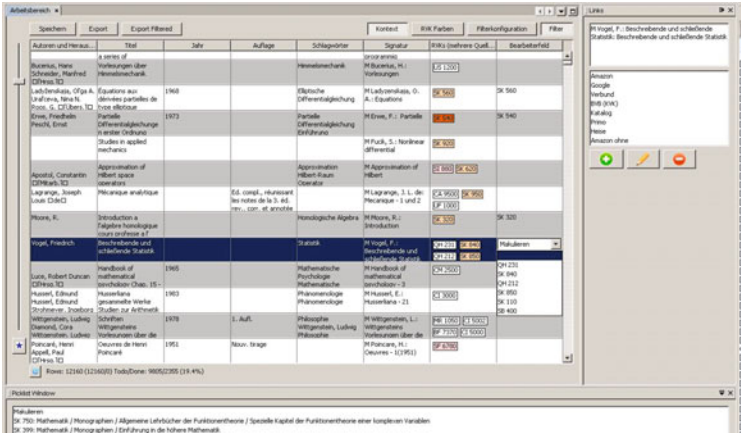
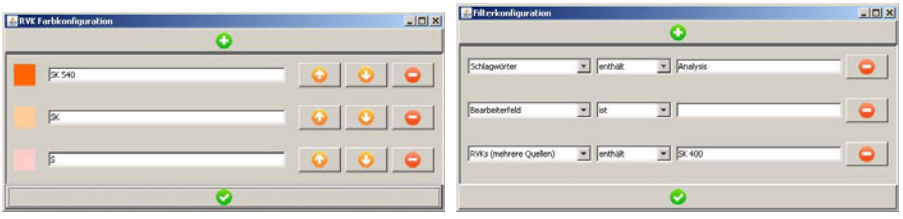


Fig. 1. The main window

Title List. Figure 1 shows a screenshot of the application's main window. Most prominent is the central tabular view of all the titles that are to be reclassified. The table view is very fast, even with large lists[2], and its contents can be sorted by any column. Columns can be hidden or shown and their order rearranged according to individual preferences and needs. In the sample screenshot the following colums are displayed: from left to right "Authors and Editors", "Title", "Year", "Edition", "Subject Headings", "Call Number", "Available Classification Numbers" and "Assigned Classification Numbers". The last field contains the decision of the user about the book represented by the table row. It can be filled in various ways:

- The user can click on one of the classes in the "Available Classification numbers" field.
- The user can select one of the values available from the drop down list. It contains the class numbers from the "Available Classifications" field, plus all recently entered class numbers.
- The user can choose a value from the *Picklist*.
- The user can assign a value to all visible rows by *Mass Assignment*.
- Finally, the user can enter arbitrary text, e.g. any class number that is not suggested or a note.

[2] Unlike common spreadsheet programs, there is no arbitrary limit of rows. We tested the application with lists containing up to 300.000 titles, and it worked flawlessly on a 3-year-old PC with 2GB of RAM.



(a) Color Configuration **(b)** Filter Configuration

Fig. 2. Adjusting the color highlighting and the filters

Throughout the application, all the class numbers can be selected to show a pop-up tooltip with the full class definition. So, the users can make themselves familiar with class numbers outside their main field of expertise.

Picklist. The *Picklist* is shown in the bottom area of the main window. The users can add any text values here to assign them quickly with a double click to the currently selected row in the table. Typical values in the picklist are a set of classes that are currently assigned to the filtered selection of books plus typical notes for actions like "Maculation" or "Presentation of the book".

Links. The *Links* window is visible on the right-hand side in the main window. It contains links to web resources that provide additional information for the currently selected title in the tabular view. The user can adapt and extend these links. Currently, per default, the following links are implemented: Amazon Book Search, Google (Books), the union catalog, a link to the local OPAC and a link to the local discovery system (Primo).

Color Highlighting. Highlighting of class numbers can be configured in various ways. As illustrated in Figure 2a, the user can chose to highlight class numbers that share a common prefix.

Filters. As shown in Figure 2b, sophisticated filtering options for all columns are available and can be combined in various ways.

3 Scenarios

In the ongoing reclassification project, we considered several common scenarios and designed the features of the application to optimally support them.

Mapping class numbers. In some cases, a class from the previously used classification system can be mapped congruently to a class from the system used for reclassification, i.e. every book with a certain class number in the old system will get the same class number in the new classification scheme. For common classification systems, such mapping sets have already been compiled [2] and can be used to speed up the reclassification process. In the application, this is implemented by first filtering the list using the "call number starts with" combination and then using *Mass Assignment* to assign the

corresponding class to all selected titles. The same steps can be used for several-to-one mappings. In the case of a one-to-several mapping, the same filter is combined with the *Picklist* to quickly assign the alternative class numbers.

Reusing available class numbers. As the classification data in the union catalog is the result of a cooperative effort of the member libraries, different codes of practice and, of course, also mistakes can be found. Therefore, the librarians at Mannheim university library feel the need to double-check the available class numbers before reusing them. By filtering the list to a single class number and highlighting potentially troublesome class prefixes, a homogenous group of titles can be produced. In this list, titles with implausible and erroneous class numbers are very conspicuous and can easily be identified for a closer inspection.

User-generated clusters. In the ongoing use of the application in our library, we noted that humans simply excel at pattern recognition. The librarians were quick to identify authors, terms or term combinations that highly correlate with a given class number and would filter and mass assign accordingly[3].

4 Experimental Results and Outlook

The application has been in use for several weeks now. As an additional feature, it collects usage statistics to help with the analysis of possible problems and to provide an insight into the usefulness of features. For each discipline a single list of titles is created from the library database and preloaded into the application for the use of the respective subject specialist. Most users started with existing mapping sets to reclassify the “easy” titles and switched to clustering the titles. In this phase of the reclassification process, we saw lasting performance rates of 600 titles/hour with peaks of 1000 titles/hour depending on the cluster size. Our librarians reported that working with the application is efficient and fun, but also tiring as there is no more dead time.

The next step in development will be the inclusion of a cutter number generator, so that the application can create complete call numbers for the reclassified titles.

References

1. Lewis, N., Seago, K.: An Automated Reclassification Project at the University of Kentucky. Cataloging & Classification Quarterly 28(4), 117–134 (2000)
2. Scott, M.L.: Conversion Tables: Set- Dewey-LC (volume 2), LC-Dewey (volume 1), Subject Headings, LC and Dewey (volume 3) (v. 1-3). Libraries Unlimited (2005)
3. Weaver, M., Stanning, M.: Reclassification project at St Martin’s College: a case study. Library Review 56(1), 61–72 (2007)
4. Werr, N., Ball, R.: Die ”neue” Regensburger Verbundklassifikation (RVK) oder die Zukunft eines Erfolgsmodells. Bibliotheksdienst 43(8/9), 845–853 (2009)

[3] As a rather unintended side effect, our librarians started to use the application to find titles that should be disposed of or moved to the closed stacks. They simply assigned ”dispose” or ”closed stacks” to the title clusters.

The Papyrus Digital Library: Discovering History in the News

A. Katifori[1], C. Nikolaou[1], M. Platakis[1], Y. Ioannidis[1], A. Tympas[1], M. Koubarakis[1], N. Sarris[2], V. Tountopoulos[2], E. Tzoannos[2], S. Bykau[3], N. Kiyavitskaya[3], C. Tsinaraki[3], and Y. Velegrakis[3]

[1] University of Athens, Greece
[2] Athens Technology Center S.A., Greece
[3] University of Trento, Italy
vivi@di.uoa.gr

Abstract. Digital archives comprise a valuable asset for effective information retrieval. In many cases, however, the special vocabulary of the archive restricts its access only to experts in the domain of the material it contains and, as a result, researchers of other disciplines or the general public cannot take full advantage of the wealth of information it offers. To this end, the Papyrus research project has worked towards a solution which makes cross-discipline search possible in digital libraries. The developed prototype showcases this approach demonstrating how we can discover history in news archives. In this demo we focus on demonstrating two of the end user tools available in the prototype, the cross-discipline search and the Papyrus browser.

Keywords: cross-discipline digital library, ontologies, keyword search, ontology browsing, multilingualism.

1 Introduction

In the last few years digital libraries have emerged providing electronic access for many user communities to information of their discipline. However, in many cases experts of one discipline turn to archives created by another discipline in the context of their research. An example of this need is the historical science, which takes advantage of archives, either cultural, scientific, press or personal, to discover information that will provide a better understanding of past events. The main problem in this process is the possible difference in the vocabulary of the historical researcher to that of the domain of the archive. This problem is related with specific challenging issues relevant to several vital research areas: coping with differences in terminology and its temporal aspects, developing techniques for semantic annotation and mappings, elaboration of query and presentation techniques for contemporary end users consuming archive information, and mitigating scalability issues. Vast amounts of digital content are available and could be incredibly useful to many user communities if it could be presented in a comprehensive to them way. The Papyrus

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 465–468, 2011.

project[1] approaches this need by introducing the concept of a Cross-Discipline Digital Library Engine. It intends to build a dynamic digital library which will understand user queries in the context of a specific discipline, look for content in a domain alien to that discipline, and return the results presented in a way useful and comprehensive to the user. To be able to achieve this, the source content has to be 'understood', which in this case means analyzed and modeled according to a domain ontology. The user query also has to be 'understood' and analyzed following a model of this different discipline. Correspondences will then have to be found between the model of the source content and the realm of the user knowledge. Finally, the results have to be presented to the users in a useful and comprehensive manner according to their own 'model of understanding'. Papyrus showcases this approach by using two domain ontologies, the history ontology as the user one and the news ontology as the content one. News archives are a major source for primary material for history researchers of different topics, ranging from political history to the history of science. This demonstration will focus on two of the tools that Papyrus offers for the end user, the Papyrus browser and the Cross-discipline search functionality, as well as on the Papyrus ontologies.

2 The Papyrus Digital Library

The conceptual flow of the Papyrus DL is depicted in Fig. 1. ***Multimedia Analysis*** includes all components that operate on the content in order to semantically annotate it with concepts of the content (news) domain ontology [5]. ***Ontology Editing and Mapping*** groups the modules which provide all the operations for building the two domain ontologies, for defining the semantic correspondences between them and for

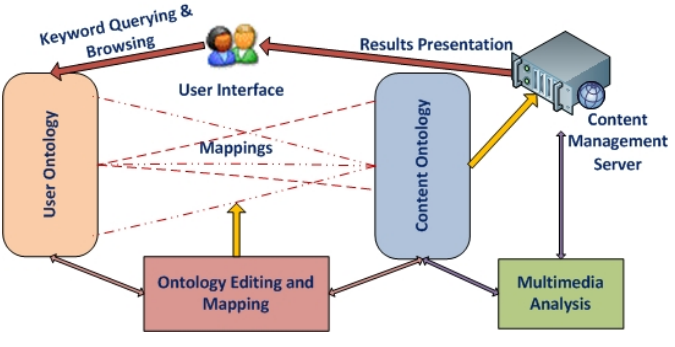


Fig. 1. The Papyrus Digital Library Engine conceptual flow

[1] FP7-ICT-215874 Papyrus Project: Cultural and historical digital libraries dynamically mined from news archives, www.ict-papyrus.eu, May 2007. The Papyrus platform was partly funded by the European Commission under the 7th Framework Programme.

semantically interpreting the user queries according to the user (history) domain ontology [4]. The ***User and Content ontologies*** [2, 3] are correspondingly history and news ontologies and the mappings provide the correspondences between them. The two ontologies have been modeled based on existing standards (CIDOC-CRM[2] and the IPTC[3] respectively) in collaboration with experts of the respective disciplines. **The *Results presentation* layer** provides the means for interfacing with the end user and accessing the underlying functionalities. ***Keyword querying and browsing*** is responsible for retrieving the information the user requests either by exploring visually the ontologies with the Papyrus browser, or by keyword search. This demonstration focuses on the two functionalities that take advantage of the history ontology to retrieve news items along with historical information: the Papyrus browser and cross-discipline search.

3 Keyword Querying and Browsing

The Papyrus end user tools to be presented in this demonstration are the Papyrus browser, a visual exploration tool that provides unified access to the two ontologies and the news content, and the Cross-discipline search functionality, which implements an ontology keyword querying technique through a visual interface.

The **Papyrus browser** [1] allows exploring news content through its association with the News ontology concepts and the corresponding mappings of these concepts to the History ontology. Besides its ability to be used as a simple Web-based ontology browser, it is a specialized tool combining two different domain ontologies and the content they describe. We will show how we can firstly select one or more historiographical issues and concepts and then retrieve news ontology concepts and related content using the mappings (Fig. 2).

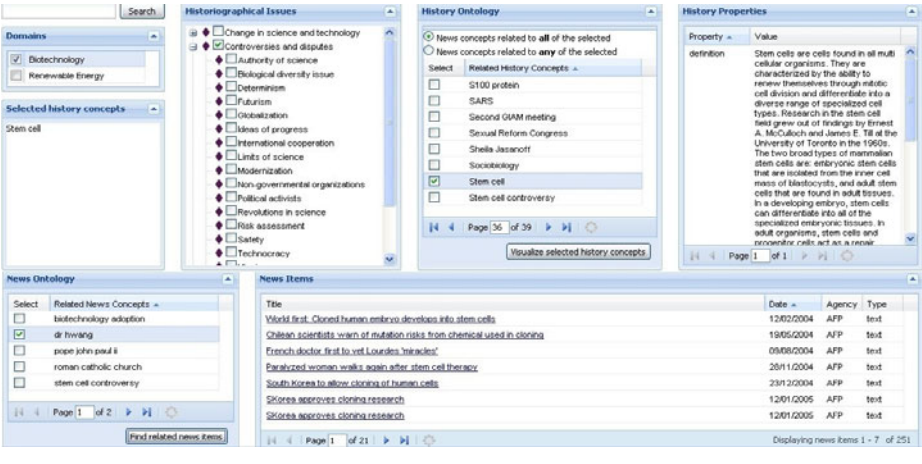


Fig. 2. Papyrus Browser– "Controversies on Stem-cells"

[2] http://www.cidoc-crm.org/
[3] http://www.iptc.org/

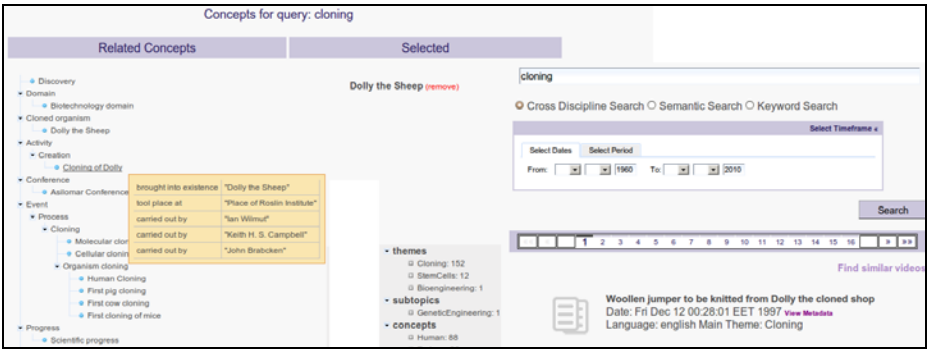


Fig. 3. Cross-discipline search - "*cloning 1960-2010*"

The **Cross-discipline search,** like the Browser, allows the user to query the History ontology, study returned History ontology entities providing the context, i.e., the secondary information related to her query, and then retrieve related news items for the selected entities. To do this, we employ an appropriate keyword search algorithm over the history ontology and the mappings between the ontologies. The query can be restricted to different time periods (Fig. 3).

4 Conclusions

The Papyrus Digital Library Engine is an integrated platform for cross-discipline search in digital archives made possible through state-of-the-art technologies. Papyrus bridges the gap between different knowledge domains and assists users in discovering information targeted to other audiences. Through the deployment of the system in the domains of history and news, Papyrus illustrates a practical example which may serve as a potential exploitable application on its own. Papyrus proves that it is possible to bridge different worlds and allow cross-discipline search through a careful indexing and mapping across their respective domains.

References

1. Platakis, M., Nikolaou, C., Katifori, A., Koubarakis, M., Ioannidis, Y.: Browsing News Archives from the Perspective of History: The Papyrus Browser Historiographical Issues View. In: WIAMIS, Desenzano del Garda, Italy (2010)
2. Kiyavitskaya, N., Katifori, A., Velegrakis, Y., Tsinaraki, C., Bykau, S., Savaidou, E., Tympas, A., Ioannidis, Y., Koubarakis, M.: Modeling and Mapping Multilingual and Historically Diverse Content. In: CIDOC, Shanghai, China (2010)
3. Kiyavitskaya, N., Katifori, Paci, G., Pedrazzi, G., Turra, R: The Papyrus News Ontology – A Semantic Web Approach to Large News Archives Metadata. In: VLDL, Glasgow, UK (2010)
4. Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: Bridging the Gap Across Heterogeneous and Semantically Diverse Content of Different Disciplines. In: FlexDBIST, Bilbao, Spain (2010)
5. Paci, G., Pedrazzi, G., Turra, R.: Wikipedia based semantic metadata annotation of audio transcripts. In: WIAMIS, Desenzano del Garda, Italy (2010)

Digitization Practice in Latvia: Achievements and Trends of Development

Liga Krumina and Baiba Holma

University of Latvia, Faculty of Social Sciences, Department of Information and Library Studies, Lomonosova Str. 1A, LV-1019, Riga, Latvia
{Liga.Krumina,Baiba.Holma}@lu.lv

Abstract. The 1980s are characterized by rapid development of digitization process and research of digital libraries. In 1994 Latvia was also involved in this process with the first attempt to digitize the materials of high demand and in poor physical condition at the Latvian Academic Library. In 1998 the digitization process was launched at the National Library of Latvia. The study, the first results of which are presented in this publication, is made to analyze the history of digitization in Latvia, and to evaluate the achievements of these activities. Up to now the development of digitization process has been poorly documented, therefore the empirical sources are unpublished documents (project reports, working papers, etc.), as well as interviews with the staff of the first projects.

Keywords: cultural heritage, digitization, digital collections, Latvian Academic Library, National Library of Latvia, memory institutions.

1 Introduction

The Latvian policy planning document "National Culture Policy (2006-2015)" gives the following conceptual conclusion of the cultural heritage:

1) "cultural heritage is the basis of sustainable development;
2) cultural heritage is a key component of identity: individual, family, group, community, region, nation, defined region of the world, e.g. Europe;
3) cultural heritage constructs people's sense of who they are, whence they come, what is meaning, value and quality of their lives" [4].

Thus, preservation of cultural heritage is necessary for the nation as a whole and the individuals who constitute it to identify their affiliation to family, community, state and maintenance of common values. Cultural heritage can be a multiform evidence of a nation-building and development.

The 21st century is characterized by a high level of development in information and communication technologies. One of ways to preserve cultural heritage is digitization that has been developed by the world's memory institutions for more than 20 years. In some cases digitization is the only way to make such material available to the future generations. The publication is a look at the digitization projects in two

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 469–472, 2011.

major libraries of Latvia (Latvian Academic Library: LAL; National Library of Latvia: NLL), and their contribution to digital cultural heritage within 1994-2006.

The problem under study is characteristic for the digitization process in Latvia. Its progress, especially at the beginning of the period, is poorly documented in materials available to public. The value, results achieved, benefits and drawbacks of the first digitization projects can be realized only in talks with the people who were involved in these activities – project managers and practitioners. The peoples' narratives, as well as the unpublished project documentation are important primary sources for description and evaluation of the beginning of digitization in Latvia. The inception of digitization identifies an important issue, as the digital objects collected and experience of digitization gained are the basis of the nascent Latvian National Digital Library (LNDL) „Letonica".

The term 'digital collection' is used to name a set of digital objects that usually include material on a particular topic by application of a simple technical solution. Materials do not have to be described and structured. This term can describe any arbitrary set of files [1].

2 Inception of Digitization Activities in Latvia

2.1 International Context

Historical research of digital libraries shows that the first digitization efforts took place at the end of the 1980s, and were aimed at digitization of scientific publications (mainly journal articles) for academic use. It was a new stage for online research and information services. The results demonstrated both technological and psychological problems and potential benefits of the use of digital documents [2]. Thirty years, which separate that time from the present, show a dynamic growth in technology, research, and volume of digital collections. Latvia has been a member of this process for the last 16 years.

In the 1990s digitization and digital collections became all-embracing in the USA, Europe, and other developed and developing countries. Formation of digital collections was a specific project based activity that involved not only IT specialists but also personnel of memory institutions in its working groups [3].

Similar trends were observed in Latvia that followed the same pattern of digitization as digitization activities in other European countries. In the case of Latvia, the first digital collections were created as a result of separate projects in two libraries – LAL and NLL. Parallel development process enabled the staff of the LAL and NLL, involved in the first projects, to learn from their European colleagues' experience in hardware and software acquisition, and workflow of digitization.

2.2 Pioneers in Latvia

Latvian Academic Library. (now – the Academic Library of the University of Latvia). Digitization in Latvia was led by the LAL. The first attempt to digitize valuable cultural materials was in *1994*, when the LAL project „Electronic library of Latvian cultural, scientific and technical information" was launched. When starting the project, digitization principles were developed for selection of materials, as well

as quality requirements were specified for master and user files. The next step (launched in *1997*) in preservation of the LAL historical stock was digitization of historian Johann Christoph Brotze's (1742-1823) collection of drawings *„Sammlung verschiedner Liefländischer Monumente ..."* in 10 volumes. Restoration of the damaged volumes was also done. Proceeding with the work of digitization, in *2000* the LAL initiated the project „Digital image archive of Baltic places". The goal of the project was to identify the poorly described images of Baltic cities and populated places in the storage units (folders, notebooks and albums) of the library. The international collaboration of the LAL started in *2001*. The successful co-operation was developed with the University of Mannheim in Germany (German poet J.M.R. Lenz's materials), and the Swiss Federal Institute of Technology Zurich in Switzerland (creation of image database). Continuing the commenced work (in *2003*), the LAL participated in the joint project „Image database of people and places" in collaboration with the Literature, Theatre and Music Museum. Within the projects the co-operation was started also with the Copyright Agency. It was successfully carried out in a form of licence, sold by the Agency for image publishing on-line. Since *2006* the LAL has participated in the creation of LNDB „Letonica". The digital collections created within the project-based activities of the LAL are available online: http://www.acadlib.lv.

National Library of Latvia. In *1998* the NLL launched a series of newspaper digitization projects. The project „Heritage-1: Latvian periodicals (1822-1940) preservation" was carried out from 1998 to 2002. Regional libraries were interested in digitization of newspapers, as they had either only partially preserved regional newspapers or did not have them at all. Museums and entrepreneurs were also involved in the implementation of the project. In *2000* the project "Posters of Latvia" was launched. The project was carried out in two stages: „Poster in Latvia, 1895-1944" and „Poster in Latvia, 1944-2000", and has resulted in a virtual gallery of posters. Continuing the commenced work (in *2002*), the NLL led the project "Latvia in the 16th-18th centuries maps". NLL maps collection reflects the historical development of the territory and geographic location of Latvia. Digital collection includes among others the works of the 16th-17th centuries cartographers and artists (e.g., Abraham Ortelius (1527-1598), Sebastian Munster (1488-1552), etc.). The digital collections created within the project-based activities of the NLL are available online: http://www.lnb.lv/en/digital-library/collections.

3 Conclusions

Digitization of Latvian cultural heritage has been going on for 16 years. In 1994 it was started by the LAL. In 1998 the NLL started digitization together with activities of other European national libraries. The analysis of digitization practice of two major Latvian libraries shows that a notable amount of digital documents was created in the form of separate collections in the early years of digitization process. A significant part of them is also available online in the form of user files. In most cases digitization was carried out within specially funded projects. The projects were financially supported by the Latvian Science Council, the Soros Foundation Latvia,

the Latvian Culture Foundation, the State Culture Capital Foundation. After the funding was stopped, the libraries continued to develop their digital collections as much as possible.

Digitization in Latvia passed three stages of development: *first stage*, the local level project based experiments (small local databases and local digital collections) in cooperation with information technology specialists; *second stage*, the state allocated targeted funds to promote co-operation among memory institutions (e. g., the state level programme „Support for co-operation among libraries, archives and museums in digital environment", 2003-2005); *third stage*, development of the unified Latvian National Digital Library „Letonica" (began in 2006), which is open to all partners for co-operation, provides the digital object management system, and will ensure the information retrieval (search aggregator). The future trends of development are linked to the mass digitization (large scale of content-rich digital objects), harmonization of metadata standards, and creation of LNDB portal. The creation of the digital library lays the foundation for uniform principles of processing, storing the digitized materials and ensuring access to them.

Assessing the early digitization activities the *value* of them can be seen in several major aspects: the clear set of objectives (why we digitize); decisions on the quality of infrastructure and imaging (how we digitize); the significant progress in documents' preservation and accessibility; the accumulated experience gained by projects' participants (particularly in the international co-operation). Several *failures* of the first projects were due to lack of knowledgeable professionals in working groups, as well as lack of experience (e.g., no experience to do test scanning using a colour scale, to set boundaries for the scanned objects, to apply metadata standards, to support the metadata harvesting, etc). The first digital projects gave the staff greater confidence and belief that together with European digitization programs the goals can be reached.

The first Latvian digitization projects focused on preservation of cultural heritage. The transfer of cultural values to digital environment and formation of a large scale digital collections will increase the interest of users (especially of younger generation) in the history of their land and their place in it. If the national digital library becomes more extensive, it will attract more users and will stimulate emotional ties with the land and the nation. It helps to realize the roots.

References

1. Dahl, M., Banerjee, K., Spalti, M.: Digital libraries: integrating content and systems. Chandos Publ., Oxford (2006)
2. Lynch, C.: Where do we go from here? D-Lib Magazine 11(7/8) (2005)
3. Tedd, L.A., Large, A.: Digital libraries: principles and practice in a global environment. K.G.Saur, München (2005)
4. Valsts kultūrpolitikas vadlīnijas (2006-2015) [State Cultural Policy (2006-2015)], http://www.km.gov.lv/lv/ministrija/vadlinijas.html

Digitizing All Dutch Books, Newspapers and Magazines - 730 Million Pages in 20 Years - Storing It, and Getting It Out There

Olaf D. Janssen

Koninklijke Bibliotheek (KB), National Library of the Netherlands,
Prins Willem-Alexanderhof 5, The Hague, The Netherlands
olaf.janssen@kb.nl

Abstract. In the next 20 years, the Dutch national library will digitize all printed publications since 1470, some 730M pages. To realize the first milestone of this ambition, KB made deals with Google and Proquest to digitize 42M pages. To allow improved storage of this mass digitization output, the KB is now replacing its operational *e-Depot* - a system for permanent digital object storage - with a new solution. To meet user demand for centralized access, KB is at the same time replacing its scattered full-text online portfolio by a *National Platform for Digital Publications,* both a content delivery platform for its mass digitization output and a national aggregator for publications. From 2011 onwards, this collaborative, open and scalable platform will be expanded with more partners, content and functionalities. The KB is also involved in setting up a Dutch cross-domain aggregator, enabling content exposure in Europeana.

Keywords: National libraries, Digital library workflows, Mass digitization, Google, Proquest, Long-term storage, Cross-domain cultural heritage, Aggregation, Interoperability, Europeana.

1 Digitizing the KB

In the early 21st century the KB started large-scale digitization of its historical publications - mainly books, newspapers & magazines - for reasons of accessibility and long-term preservation. The focus so far has been on creating full-text corpora for study and research in the humanities using **public** funding. Results by the end of 2011 will include 2.3M pages of *Dutch Parliamentary Papers* [1], 8M pages from popular Dutch regional, national and colonial *Historical Newspapers* [2], 2.1M full-text pages of *Early Dutch Books* [3] and 1.5M pages from frequently consulted old magazines.

[1] Filming and digitization of the Dutch parliamentary papers 1814-1995, http://www.kb.nl/hrd/digitalisering/archief/staten-generaal-en.html (project information) & http://www.statengeneraaldigitaal.nl/ (website)

[2] Dutch Historical Newspapers 1618-1945, http://www.kb.nl/hrd/digi/ddd/index-en.html (project information) & http://kranten.kb.nl (website)

[3] EDBO – Early Dutch Books Online - 10.000 full-text digitized books from 1781-1800, 2.1 million pages, http://www.earlydutchbooksonline.nl

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 473–476, 2011.

In 2010 the KB announced its ambitious plans to digitize all Dutch books, newspapers, magazines and other printed publications from 1470 onwards, a total of 730M. A first milestone is set for 2013, by when the library should have scanned 10% of this amount. To realize its ambition, the KB cannot not rely on public funding alone. It has therefore entered into strategic **public-private** partnerships with both Google[4] and Proquest [5] to digitize 210.000 books (some 42M pages) from its public domain collections.

2 Permanent Storage, Now and in the Future

As the national library, the KB has a responsibility to permanently store not only printed publications, but also digital manifestations. To perform this duty the KB joined forces with IBM in 2000 to build the world's first OAIS-based processing and preservation system for long-term storage of digital objects. In 2003 this resulted in the operational *e-Depot* [6]. Nowadays, this deposit is a safehaven for over 15M scientific articles from some of the world's biggest STM publishers [7].
In 2012 the KB's maintenance contract with IBM will run out and components of the system will no longer be supported. With the 'seven-year-itch' or the system [8] having past, it is already living longer than most other IT systems. Further reasons for upgrading the e-Depot include

- *Volume & scalability:* the KB wants to permanently store the hundreds of millions of files resulting from its mass digitization programme output.
- *Heterogeneity & flexibility:* the current system is optimized for processing and storing relatively small numbers of homogeneous single objects (PDFs). Giving fast access to large numbers of diverse and compound content (such as e-books) will become increasingly common in the near future.

The KB is a partner in the *SCAPE* project [9]. This initiative will provide ongoing technical innovation by developing scalable preservation planning and execution services that can be deployed in the new e-Depot system within the next 3-5 years.

---

[4] KB and Google sign book digitization agreement, http://www.kb.nl/nieuws/2010/google-en.html

[5] Digitization by Proquest of early printed books in KB collection, http://www.kb.nl/nieuws/2011/proquest-en.html

[6] E-Depot, the KB's digital archiving environment for permanent access to digital objects - http://www.kb.nl/hrd/dd/index-en.html

[7] Including, but not limited to Elsevier, BioMed Central, Blackwell Publishing, Oxford University Press, Springer and Brill. For a complete list, see http://www.kb.nl/dnp/e-depot/operational/background/policy_archiving_agreements-en.html

[8] Wijngaarden, H. van.: The seven year itch. Developing a next generation e-Depot at the KB. Paper for the 76th IFLA General Conference and Assembly, 10-15 August 2010, Gothenburg, Sweden, http://www.ifla.org/files/hq/papers/ifla76/157-wijngaarden-en.pdf (accessed on 28-03-2011)

[9] SCAPE - SCAlable Preservation Environments, http://www.scape-project.eu/

3 Providing Access

The back-end data standards [10] are identical across all KB-run mass digitization projects, making the outputs in theory fully interoperable. However, this potential has not yet been optimized in the front-end presentation of the KB's full-text collections. So far this has been done via separate, websites (1,2,3,[11]), each with its own specific branding, URLs, design and search & object display functionalities. For end-users the KB-collections thus appear to be unrelated and scattered, making them relatively difficult to use given their demands.

The KB has taken the needs of its users seriously and has just finished designing and implementing the first basic iteration of the *Dutch National Platform for Digital Publications* (working name). This full-text content distribution platform will give access to digitized books, newspapers and magazines, including the output of the KB's mass digitization projects. Access will be central via a modern web2.0 site, as well as distributed via search and display APIs for delivering content to the places and networks the user are.

Furthermore, it will be positioned as a full-text and metadata aggregator, with the aim of making the content interoperable and exporting it to cross-domain initiatives, both on national, European and global levels.

4 The Cross-Domain and International Dimensions

As the national library, the KB has a very important facilitating and networking role in the Dutch scientific and cultural infrastructure. Using this position, it has the potential to set up and stimulate different levels of collaboration to make online heritage more accessible. This is illustrated by the 3-tier collaborative model in Fig.1.

Lower level: domain specific collaboration & aggregation
As said in Section 3, KB's *National Platform for Digital Publications* will be positioned as an aggregator for Dutch full-texts, aiming to make the content - and the network of content delivering partners - interoperable and ready for participation in cross-domain initiatives on national and international levels. Similarly, organizations from other domains are working on interoperability and aggregation for their specific sectors, as indicated by the spheres on the lower level.

Middle level: national cross-domain collaboration & aggregation
To enable these sector specific aggregation initiatives to come together, the results of the *NED!* project [12] are used. It delivered a basic infrastructure for the interoperability of Dutch digital heritage, using open standards including XML, DublinCore, OAI-PMH and SRU. It is now being expanded to build a cross-domain heritage aggregator that can become *the* national hub for content delivery to international initiatives.

---

[10] KB's open digitization & accessibility standards, http://www.kb.nl/hrd/digitalisering/standaarden-en.html

[11] Digitization of ANP news items, http://www.kb.nl/hrd/digitalisering/archief/anp-en.html (project information) & http://anp.kb.nl (website)

[12] NED! - Nederlands Erfgoed Digitaal!, http://www.nederlandserfgoeddigitaal.nl/

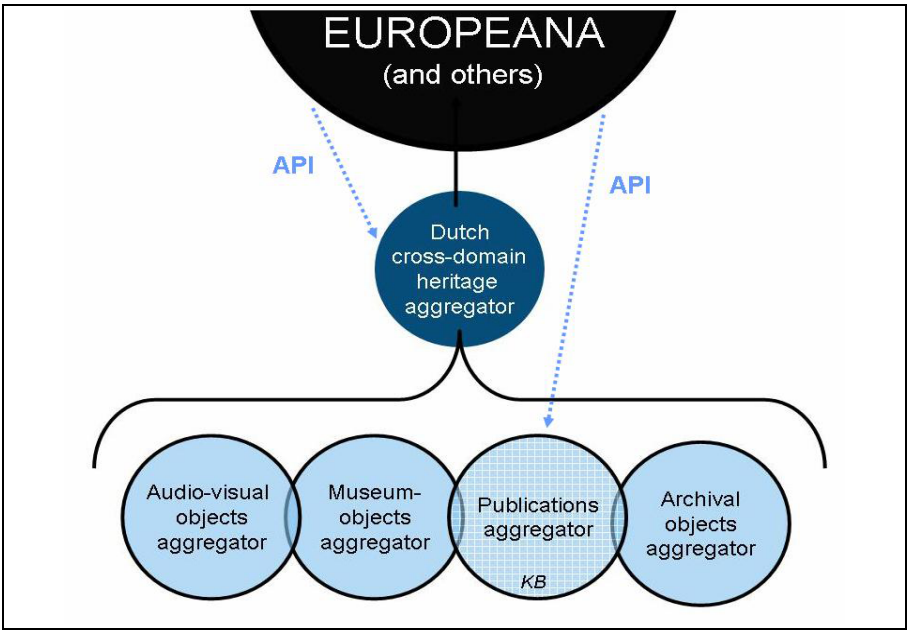


Fig. 1. Dutch national collaborative aggregation model. The KB is responsible for aggregating publications in the *National Platform for Digital Publications.*

Building a national aggregator is however a step-by-step process, not finished overnight. Until that time domain-specific aggregators - in case of the library domain the *Dutch National Platform for Digital Publications* or *The European Library* [13] - will continue to have an important role in routing Dutch library content directly to top-level services.

Top level: International cross-country collaboration & aggregation
Having established national cross-domain aggregation and interoperability on as many levels as possible [14], Dutch content can be shown and used on international stages, most notably Europeana [15]. This service offers a set of APIs [16], which not only enable reuse of Europeana content by third parties, but also allow the contextualized and enriched content of the providing institutions to be used in their own environments. The APIs, in other words, make it possible to create user interface elements for (dark) aggregation services on the lower and middle levels, as indicated in Figure 2 by the dotted API arrows.

[13] The European Library; on the one hand a free service that offers access to the resources of the 48 national libraries of Europe in 35 languages, on the other hand an international library domain aggregator for Europeana, http://www.theeuropeanlibrary.org
[14] Establishing interoperability on as many levels as possible: technical, metadata, semantical, human, inter-domain, organizational, political, .etc.
[15] Europeana; paintings, music, films and books from over 1500 of Europe's galleries, libraries, archives and museums, http://www.europeana.eu
[16] Europeana Application Programming Interfaces, http://version1.europeana.eu/web/api

Design, Implementation and Evaluation of a User Generated Content Service for Europeana

Nicola Aloia[1], Cesare Concordia[1], Anne Marie van Gerwen[2], Preben Hansen[3], Micke Kuwahara[4], Anh Tuan Ly[5], Carlo Meghini[1], Nicolas Spyratos[5], Tsuyoshi Sugibuchi[5], Yuzuru Tanaka[4], Jitao Yang[5], and Nicola Zeni[1]

[1] Istituto della Scienza e delle Tecnologie della Informazione, National Research Council, Pisa, Italy
[2] EDL Office, The Hague, The Netherlands
[3] Swedish Institute of Computer Science, Sweden
[4] MEME Media Lab., University of Hokkaido, Sapporo, Japan
[5] Laboratoire de Recherche in Informatique, University of Paris South, Orsay, France

Abstract. The paper presents an overview of the user generated content service that the ASSETS Best Practice Network is designing, implementing and evaluating with the user for Europeana, the European digital library. The service will allow Europeana users to contribute to the contents of the digital library in several different ways, such as uploading simple media objects along with their descriptions, annotating existing objects, or enriching existing descriptions. The user and the system requirements are outlined first, and used to derive the basic principles underlying the service. A conceptual model of the entities required for the realization of the service and a general sketch of the system architecture are also given, and used to illustrate the basic workflow of some important operations. The planning of the user evaluation is finally presented, aimed at validating the service before making it available to the final users.

Keywords*:* User Generated Content.

1 Introduction

In the 2011-2015 Strategic Plan, Europeana [1] announces User Engagement to be one of the strategic tracks by which the organization will deliver value. By the term 'Engage' Europeana refers to cultivating new ways for end user to participate in their cultural heritage. The Europeana network comprises communities of archivists, curators and librarians who show a growing interest in exploring new methods of access and dialogue. Europeana intends to enhance the user experience and offer services that allow users to interact and participate.

User-generated-Content (UGC) is one aspect of this renewed way of participating. Information about cultural heritage exists outside the heritage institutions; artifacts, written sources and memories of individuals complement collections held in institutions. UGC services are designed to provide users with means to support and interpret content. They will be involved in storytelling, curating of virtual exhibitions, reviews and even the creation of new collections. Greater participation will increase

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 477–482, 2011.

users' interest and loyalty. Europeana is therefore devoting increasing resources to initiatives that bring out the value of the contribution those users can make. In response to these needs, the ASSETS [2]. Consortium has included the support of user-generated content amongst the services it is going to develop for Europeana. ASSETS is a two-year Best Practice Network co-funded by the CIP PSP Programme to improve the accessibility and usability of Europeana. Rather than focusing on a specific set of UGC applications, ASSETS is developing a general purpose, back end component that aims at supporting any UGC service Europeana will want to offer to its users. To this end, the ASSETS back end component implements an Application Programming Interface (API) for creating, storing and manipulating UGC Units of Work, and for submitting these Units of Work to Europeana, in the form of Europeana Submission Information Packages (SIPs). Final users will interact with their Units of Work through client interfaces, which will hide the unnecessary technical details and complexities of the back end to them, providing them with the level of representation that is most suitable for the specific UGC task at hand. Indeed, it is expected that every UGC task will be supported by a different final user interface. But this will have no impact on Europeana, since every different front end will talk to Europeana through the same API. The API will relieve future UGC applications from implementing any server side functionality and will move away from Europeana the technical interoperability problems that would arise upon integrating into its database the possibly different objects coming from future UGC applications. The service will rely on the Europeana Data Model (being developed by the Europeana version 1.0 project [3]) in order to tackle the more serious semantic interoperability problems.

The definition of the conceptual model underlying the UGC API is the most difficult challenge that the ASSETS UGC team is facing. The model has to strike the optimal balance between simplicity, so to be quickly learned and easily coded against by the future UGC service developers, and generality, so to satisfy the needs of any possible future UGC service. This conceptual model has been defined during the first year of the ASSETS project, based on an analysis of the different types of requirements that are in place. The model has been subsequently used to define the UGC API.

2 Requirements

User requirements can take different forms: (a) Submission of objects to a repository. A minimal set of metadata will have to be provided in order to support the interpretation, discovery and management of the object. The user requires to be free of choosing which metadata format to use, but the system must propose a default one. (b) Metadata enrichment: Users contribute factual metadata to an object, such as location, date, names, or tags. The object can be created by the user, but also by another user; the object may also be existing content in Europeana. (c) Annotations: users are contributing their views, comments. opinions to an object. (d) Contextualization: through storytelling or creating virtual exhibitions and galleries and possibly adding narratives to them, users are combining existing objects into a new context (without changing the objects and metadata itself). Before publishing user generated content, moderation may be added as

an intermediate step in the process. Authorized users review the UGC and decide to accept and publish it. In most cases, this includes a feedback loop to the user who originally contributed the data.

Europeana audiences include academic researchers with a high level of language and computer skills but also people who are hardly familiar with foreign languages or using the internet. While possessing intermediate to good knowledge of foreign languages and online search, these groups generally expect services to be easy and intuitive. At the same time, they want to understand what happens with their contribution and who keeps control over their content. User Interfaces should therefore preferably be simple, straightforward and visual. Additionally, clear information must be provided about rights regarding the content.

The back end component has to comply with the Europeana architecture, which is based on an Open Source policy. Europeana has also defined a set of guidelines [4] regarding the coding, the testing and the deployment of the components that make up its architecture.

3 The Conceptual Model of the UGC Service

In order to meet the user and system requirements, the ASSETS team designed a UGC service based on the concepts outlined below and presented in Fig. 1 as a UML class diagram.

From the UGC server point of view, at any point in time there exists a set of users of the UGC service. Each user is in fact a role, identified by an id and a password, behind which a whole community may actually operate. Each user has its own Workspace (WS) on the UGC server. A WS is simply a container of the objects that the associated user needs to perform UGC tasks.

The creation of a single UGC object may take a long time and span several sessions of work. In between one of these sessions and the next, the partial results achieved so far have to be persisted, in order not to be lost and to be resumed at the beginning of the next session. The concept of "partial" UGC is captured by the notion of Unit of Work (UoW). The UoWs of a user are maintained in the user's WS.

A single UoW contains objects, identified by URIs, and their accompanying descriptions. The objects in a UoW can be of two kinds:

- Existing Europeana objects, that the user has included in the UoW in order to link them to new objects (see below) as values of some property, or in order to enrich them with new descriptions. Existing Europeana objects can be retrieved for inclusion in a UoW via a query issued to Europeana.
- Newly created objects, which are called UGC objects. These objects are original contributions to Europeana, and can be of three kinds:
    - digital objects having an associated media file with the content;
    - digital objects for which no media file is available;
    - non-digital objects.

Every object in a UoW has an associated description. A description represents a metadata record of the object and is modelled as a set of attributions, each attribution consisting of a property and a value. Different attributions can have the same property with a different

value. A value can be itself an object, or a literal or another resource, external to the digital library.When a UoW is ready to be submitted to Europeana, the user can do so by using an operation that transforms the UoW into a well-formed Submission Information Package (SIP) and places a message signalling the existence of the SIP into the Outbox. Each user WS is endowed with an Outbox. Europeana retrieves messages from Outboxes in order to harvest the corresponding SIPs. As already mentioned, users can issue queries to Europeana in order to retrieve objects. Each query returns a result, in the form of a message stored in a special area of the user WS called the Inbox. Each user WS is endowed with an Inbox. Messages in the Inbox are of two kinds: query results and notifications that communicate the result of submissions. In case of a negative notification the rejected SIP can be retrieved and re-transformed into a UoW so to allow the user to perform the necessary repairing actions. It is important to notice that these concepts define a general-purpose schema, whose machinery need not be used by every UGC application. For instance, a simple UGC task that takes place in a single session, such as an image upload, may be implemented by directly building the corresponding SIP, so by-passing the UoW stage. On the other hand, another UGC application may decide to publish a finished UoW to a community of users in order to perform a socially oriented form of mediation before submitting the UoW to Europeana. These decisions will be taken by the client side of the applications, relying on appropriate shortcuts offered by the UGC API.

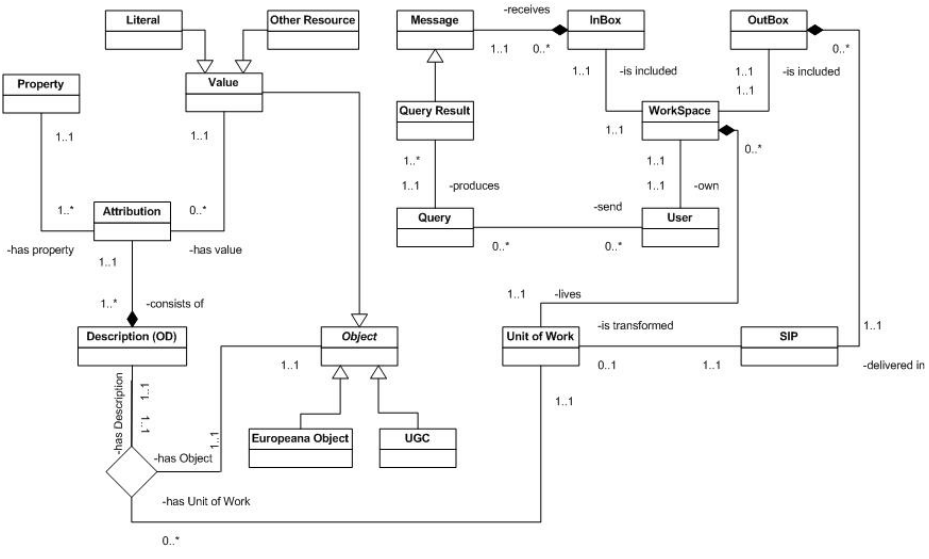


Fig. 1. UGC model description

4 Architecture

For the purposes of developing, testing and evaluating with users the UGC functionality, the UGC server will be deployed on the ASSETS Server. After successful evaluation, the Server will be moved into the Europeana production server.

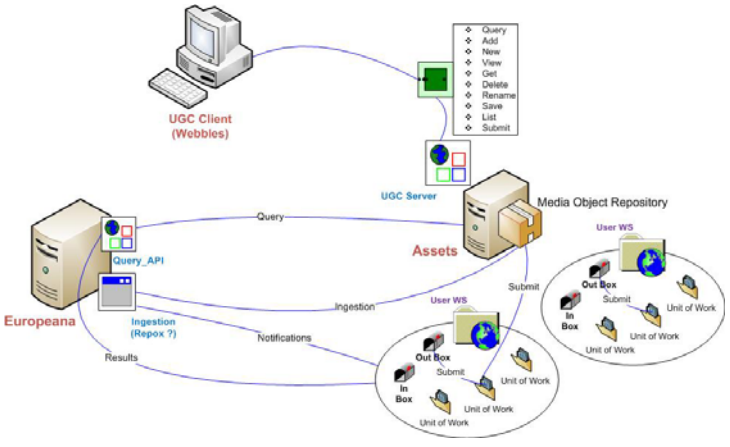


Fig. 2. Overall architecture

The three main components of this architecture (see Fig. 2) are:

Europeana Server. This is the server implementing the Europeana search functionality. It provides an API to search for content in the digital library, it implements the Open Search directives [5]. The Europeana server also provides an application for harvesting of data.

ASSETS UGC Server. In the context of UGC, the ASSETS UGC server provides functionality to communicate with Europeana and with the UGC client. The communication with the Europeana server is by invoking the Query API module (namely OpenSearch API). The ASSETS UGC server manages the workspaces of the users, and provides an API to manipulate the UoWs in the workspace. The UGC server provides API as REST Web services and is independent from any specific UGC client. The User WS contains an Inbox, an Outbox and the set of Units of Work that the user is currently playing with. In addition, the UGC Server maintains the ASSETS Media Object Repository (AMOR), implementing OAI-PMH functionality to allow Europeana to harvest SIPs.

UGC Client. Is a browser-based GUI supporting the user in a specific content generation tasks. The UGC Client interacts with the ASSETS Server via REST web services provided by the UGC Server module.

5 Evaluation

Digital Libraries, Archives and Museums represent the main target for Europeana and through them it will be possible to gain a better understanding of user needs and requirements that will contribute to the design of future services for the users in general as well as for specific groups of users and within specific environments such as the mobile usage and social media applications.

Within the Europeana project, there will be two types of evaluations carried out: a technical and a user-centred evaluation. The technical evaluation will be focused on verifying that the functionalities delivered by ASSETS fulfil efficient and scalable requirements. The user-centred evaluations aim at verifying that the services delivered by ASSETS fulfil the expectations of the end-users. The EDL Foundation together with ASSETS will conduct the evaluation. More specifically, the goal for the evaluation will be a) to utilize a user-oriented approach, and b) to focus on usability aspects of the services proposed and their end-users.

6 Conclusions and Outlook

The main concepts and architectural features of the user-generated content service have been illustrated. The service is being implemented by the ASSETS Best Practice Network and will be evaluated within the lifetime of the project. The basic principles of the evaluation methodology have been described.

The UGC service developed by ASSETS is based on a general-purpose back end, which is meant to relieve Europeana from dealing with the specificities of the possibly very many UGC tasks that may be offered to users. At the same time, the back end relieves developers of UGC tasks from implementing the server side of their applications.

Acknowledgements. This work has been partially supported by the PSP Best Practice Network ASSETS (CIP-ICT PSP-2009-3, Grant Agreement n. 250527).

References

1. The Europeana portal, `http://www.europeana.eu`
2. The ASSETS project, `http://www.assets4europeana.eu`
3. The Europeana version 1.0 project, `http://version1.europeana.eu`
4. The Europeana Labs, `http://europeanalabs.eu/wiki/`
5. The Open Search, `http://www.opensearch.org/Home`

Connecting Repositories in the Open Access Domain Using Text Mining and Semantic Data

Petr Knoth, Vojtech Robotka, and Zdenek Zdrahal

Knowledge Media Institute, The Open University
United Kingdom
{p.knoth,v.robotka,z.zdrahal}@open.ac.uk

Abstract. This paper presents CORE (COnnecting REpositories), a system that aims to facilitate the access and navigation across scientific papers stored in Open Access repositories. This is being achieved by harvesting metadata and full-text content from Open Access repositories, by applying text mining techniques to discover semanticly related articles and by representing and exposing these relations as Linked Data. The information about associations between articles expressed in an interoperable format will enable the emergence of a wide range of applications. The potential of CORE can be demonstrated on two use-cases: (1) Improving the the navigation capabilities of digital libraries by the means of a CORE pluging, (2) Providing access to digital content from smart phones and tablet devices by the means of the CORE Mobile application.

Keywords: digital library federations, automatic link generation, text mining, semantic similarity, content harvesting, mobile devices.

1 Introduction

In this paper, we present CORE, a system which aims to improve the access and navigation between semantically related Open Access articles stored across different repositories. Our approach employs text mining to discover relations between full-text content. These relationships are then represented and published as Linked Data. We have developed two software applications that demonstrate the use of this new dataset: the CORE Plugin and the CORE Mobile application.

Generating and exposing information about semantically similar papers using CORE involves three stages. In the **first stage**, the metadata records and full-text documents are harvested from available Open Access repositories (Metadata Harvester and Full-text Downloader components - Figure 1). In the **second stage**, the available content is processed in order to discover meaningful relations between papers using text mining techniques (Semantic Relation Analyser component - Figure 1). In the **third stage**, the extracted relations are exposed as Linked Data and are represented in the CORE Triple store. Each stage is now described in more detail.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 483–487, 2011.

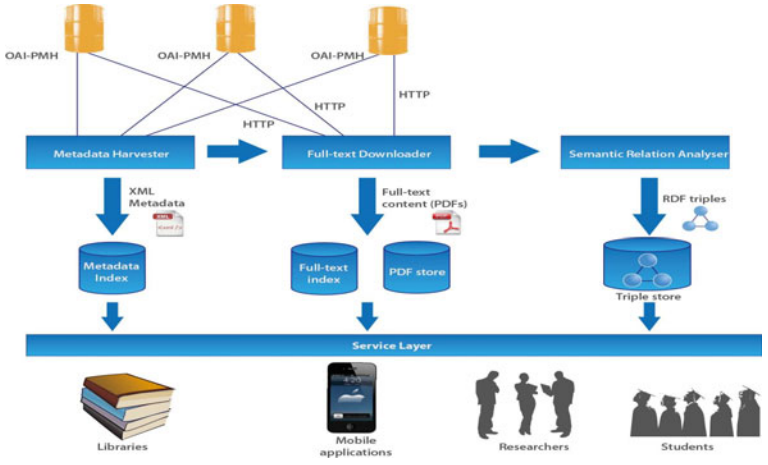


Fig. 1. CORE architecture

2 Metadata and Content Harvesting

The first component of the CORE system is responsible for acquiring (1) metadata records and (2) the associated full-text content from Open Access repositories. The harvesting of the metadata is performed using standard OAI-PMH requests to the repositories. Successful requests return an XML document which contains information about the papers stored in a repository. Although the OAI-PMH protocol itself is not directly concerned with the downloading of full-text content, as it focuses on the transfer of metadata, a good practise in repositories (which is unfortunately not consistently applied) is to provide as a part of the metadata the URLs to the full-text documents. Document URLs can be thus extracted and used to automatically download full-texts from repositories over the HTTP protocol. The CORE system provides this functionality and is optimized for regular metadata harvesting and full-text downloading of large amounts of content. The fact that CORE caches the actual full-text content in order to process the documents and to discover additional metadata distinguishes this approach from a number of other Open Access federated search systems, such as BASE or OAISTER, that rely only on the metadata accessible through OAI-PMH.

At the time of writing, the CORE harvesting system has been tested on 142 Open Access repositories from the UK. We expect to extend the number of repositories in the future to all available repositories listed in OpenDOAR. A larger number of repositories will increase the demand on the storage space as in our case not only metadata, but also full-text has to be stored as opposed to services, such as OAIster or BASE. It is expected that about 10% of the records contain links to downloadable pdfs, which would account for about 3 million full-text papers stored across Open Access repositories worldwide.

3 Discovering Related Content

The analyzer of the semantic relations from textual content is at the heart of the CORE application. This component performs the extraction of text from the downloaded papers and then processes the content by calculating the semantic similarity between pairs of articles. The current implementation used in this calculation is an adapted version of software, which we have originally developed to study the correlation between links authored by people and links predicted by automatic link generation methods, namely using semantic similarity measures on document vectors extracted from text [2]. This study revealed that semantic similarity is strongly correlated to the way people link content and that the value of the calculated semantic similarity can be used to predict useful connections between articles while filtering out duplicities.

The system is based on the calculation of cosine similarity between $tfidf$ vectors. Since calculating semantic similarity for all document pairs doesn't scale up, due to the large number of combinations, CORE uses heuristics to decide which document pairs are unlikely to be similar and can thus be discarded from consideration.

4 Exposing Semantic Data

In the third stage, the newly generated metadata about the article similarities are expressed in an interoperable format. This should allow third-party services to exploit this information in flexible and unprecedented ways. When exposing a new collection as Linked Data, it is a good practise to reuse existing vocabularies/ontologies for its description as this makes it easier for the outside world to integrate the new data with already existing datasets and services. The Similarity Ontology - MuSim [1], has been selected as an appropriate base schema for the representation of relationships between papers within CORE. Though this ontology was originally designed with music similarity in mind, it can be easily applied to other domains. In MuSim, the association between two (or more)

Fig. 2. (a) Representing similarity between articles using MuSim. (b) CORE Mobile application.

top level objects is a class to be reified rather than a property. This allows to embrace the complexity of associations and accommodate the subjectivity and context-dependence of similarity. Figure 2 shows how information about similar items is represented in the CORE system. Using this representation it is possible to construct complex queries combining information from various datasets.

5 CORE Services

The CORE Linked Data set can be reused by third-party applications in a number of ways. We have developed two applications that demonstrate its use.

(1) The CORE Plugin - The CORE Plugin can provide information to digital libraries about related documents stored in other repositories that are semantically related to to the document currently being visited. This provides the following advantages to the libraries: (a) The repository/library content is compared to external content remotely using the CORE system eliminating the needs for additional memory storage or computational power. This means that all the functionality is provided to third-party systems via a web-service. The web-service communicates with the client application which allows flexible integration with any repository user interface. (b) By linking to relevant information stored within this repository and elsewhere, the repository/library provides better service to its users.

(2) CORE Mobile - Mobile devices are often used in situations when network access is limited. Therefore, it is practical to be able to access the content of interest also in an off-line mode. CORE Mobile is a native application (Figure 2) for the Android system, which can be used on both mobile and tablet devices. The application allows searching and navigating across related papers stored in Open Access repositories. The client application connects to the CORE service layer through which the search for relevant articles and the successive browsing across related content is available. The application also allows the downloading of full-text content to the mobile device and the accessing of the downloaded content when off-line. The application is freely available from the Android Market.

6 Conclusion

This paper presented the CORE application, which automatically discovers related content across Open Access repositories. The applied approach is fully automatic and relies only on the harvesting of metadata and content using existing protocols. CORE is to our knowledge the first system which discovers relations between papers based on the processing of full-text information as opposed to existing metasearch systems. In addition, the information about associations between papers, harvested from different digital repositories, is exposed in a format defined by a similarity ontology and published as Linked Data, making it possible for third-party systems to exploit the potential of the CORE dataset. Finally,

we have demonstrated two cases in which the CORE dataset has been used: (a) as a plugin into existing digital library system (b) as a federated search system with download and browsing capabilities for mobile devices.

References

1. Jacobson, K., Raimond, Y., Gangler, T.: The similarity ontology - musim (2010)
2. Knoth, P., Novotny, J., Zdrahal, Z.: Automatic generation of inter-passage links based on semantic similarity. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, China, pp. 590–598 (August 2010)

CloudBooks: An Infrastructure for Reading on Multiple Devices

Jennifer Pearson and George Buchanan

FIT Lab, Swansea University
{j.pearson,g.r.buchanan}@swan.ac.uk

Abstract. The use of light, portable devices such as iPads whose reading angle is readily changed is radically different to reading on a desktop or laptop. However, it would be naive to view this as mere evolution. Rather, such devices permit reading activity to more closely mirror paper. A light, keyboardless device can be used in many different locations and orientations. This paper reports an infrastructure for supporting reading on multiple slate devices using a single cloud-based system to provide for numerous configurations.

Keywords: Slate PCs, Collaboration, Digital Reading, Annotation.

1 Introduction

The central role of slate PC devices such as the iPad is "media consumption", including the reading of electronic books. Superficially, this simply transfers existing software from the desktop (or laptop) PC onto a new form factor. Apps such as GoodReader and iAnnotate primarily replicate the interaction design found in Adobe Acrobat and other desktop document reading applications. In this paper, we introduce a basic infrastructure, CloudBooks, that supports multiple reading devices (primarily, slate PCs), which may be connected in co-located or remote, synchronous or asynchronous, single- or multi-user configurations.

By placing key features in a network-based 'cloud' infrastructure, the CloudBooks services become ubiquitous: e.g. mark-up can be transferred and communicated quickly between multiple devices. Previous systems such as Polar [4] and DiLAS [1] have demonstrated the viability of providing an extensive annotation function which established the value of providing an external service that integrates with a central DL server. There is, in short, an established corpus for separating annotations from the document itself, using some form of continuous annotation service, that serves as a repository for the notes applied to documents by the library's users. CloudBooks builds on these existing principles and adapts them to a user's own, informal collection rather than a DL.

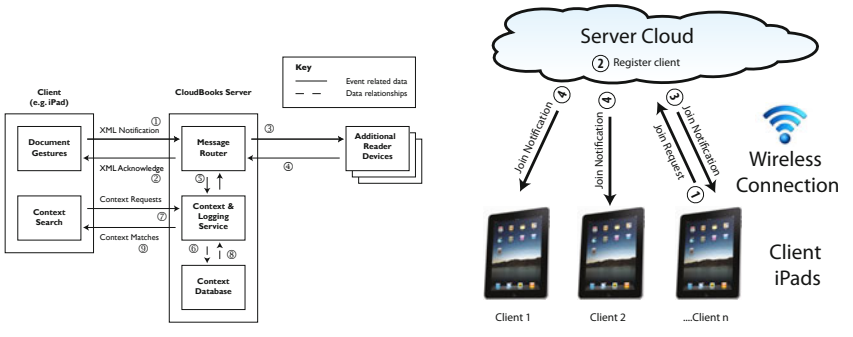
2 CloudBooks Architecture

The general CloudBooks Architecture is shown in Fig 1a. On the left appears a single iPad (or slate) device; further devices are depicted on the right. In the centre appears the CloudBooks server, and its associated components.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 488–492, 2011.

To connect to the server, an iPad must first register with the message router component using its IP address as an identifier. The message router is later used to forward messages received from one device to another. Each registered device belongs to one (or more) groups, and a message is normally forwarded through the message router to all members of the group. To take a concrete example, when a device wishes to send an event to its group, it first forwards an XML message to the router ①; the router sends a response to confirm its receipt of the message ② and immediately forwards the message to the other members of the group ③, which also send acknowledgements of receipt ④. Optionally, messages can be saved to the log service ⑤ and its database ⑥.

The communication support provided by the message router provides no long-term storage, and simply provides the same messaging capability as the Greenstone Alerting Service [3]. CloudBooks can store contextual data using its logging components (centre, middle and bottom). This can be achieved through the log service (steps ⑤ and ⑥). This service can also be used directly ⑦; either to retrieve content or history (Fig 1a), or for logging. When contextual information from the log service is needed, a request can be sent ⑦, requesting ⑥ and retrieving ⑧ content that is then returned to the client ⑨.



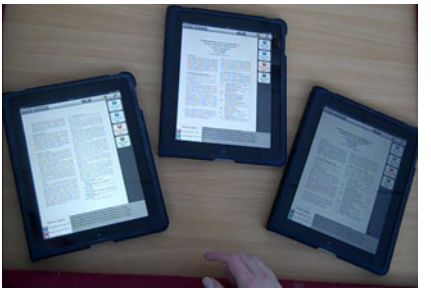
(a) The CloudBooks Architecture **(b)** The gsBooks Architecture

Fig. 1. CloudBooks and gsBooks Architectures

3 Reading in a Group: Multiple Devices, Multiple Users

The act of reading is typically solitary, e.g. reading a newspaper. However, there are occasions where reading is performed interactively in groups, e.g. in study groups, where there are several ways of reading a document: single paper copy, multiple paper copies, single computer, multiple computers, single large display. Each method has its own strengths and weaknesses [2]. Paper is easily manipulated and does little to impede interaction between people; readers can sit wherever there is space without technological constraints. A single paper copy of a document (or indeed a single computer) can cause space and 'group leader' problems whereas multiple paper copies can be tricky to reference.

(a) gsBooks Interface



(b) gsBooks Collaboration

Fig. 2. gsBooks

To better support collaborative reading, we have implemented a subset of CloudBooks called gsBooks; a system that makes use of multiple iPads to facilitate real-time communication between multiple users. The device's form factor allows users to sit around the same desk while also giving them a personal document view. The iPad's WiFi allows users to contribute easily via real-time changes to all iPads in the group. This not only facilitates simultaneous annotation between group members, but also raises user awareness of the notes made within the session. It supports both local and remote collaboration via our cloud–based infrastructure.

The gsBooks architecture is shown in Fig 1b. To connect to a group, a client iPad sends an XML join request ①. CloudBooks then registers the user ②: allocating a unique colour[1] to the client, and adds the client IP address and nickname to the active client list. Following a successful join, the server then distributes a join confirmation back to the client ③ and then to all other clients within the group ④. When an action (see below) is made by a client, it sends a new XML 'action' to CloudBooks. The server updates its own list of the annotations on the document before distributing the change to all active clients. By keeping its own complete copy of the annotations, CloudBooks can supply new or returning clients with a complete history of the document.

The gsBooks system supports three main document activities: annotation, bookmarking which are permanent user notes, and 'point outs' which are temporary markers that coordinate the reading of group members. When reading in a group, users need to indicate content to other members, e.g."look at this figure on the right of page 45". When working on a single document this process is straightforward: the user can physically point to the section. This problem is more difficult when each user has their own copy of the text, particularly if there are many users or if group members are sitting far apart. To aid in this process then, we have implemented the 'look at this queue' (shown in Fig 2a on the left of the screen): a tool that allows users to quickly point out specific sections of the

[1] For consistency, users that disconnect from the working group, then later re-connect will be assigned their original colour.

working document. When a point out is made, a new entry which includes the nickname and colour of its creator, is added to the top of every group member's 'look at this queue'. Clicking on this entry then takes the user directly to the page and points out the exact area the point out was made.

4 Reading Alone: Multiple Devices, One User

Cloudbooks can also capture reading behaviours over a significant span of time, so a user's reading history is enriched by recording details of what was read on different devices or locations. We have developed another subset of CloudBooks, called xBooks, that captures details of a reader's history (e.g. when a document was opened) which provides a unified log of the user's reading behaviour. If such a log is maintained on a single device, then under most circumstances that information cannot be retrieved remotely. However, a variety of options are made available when using CloudBooks. One configuration stores a log of the reading on each copy of xBooks within the CloudBooks service. A user can then query what they read, or when they read it. Adding contextual information such as GPS (when available) can permit location to be inferred from that data. Hence, one can (at least approximately) retrieve what one read at a given place.

The first prototype for single users provides the ability to distribute reading histories across multiple devices, building on the basic configuration just described. This prototype mirrors the functionality available of the Kindle - i.e., reading history across *all* reading devices are synced. However, unlike the functionality provided by Amazon, we support both local and global histories. A user can search against their history using the location, device, time or title.

The second prototype uses CloudBooks to control the view presented on multiple iPads via a single document reader interface. The iPad views can thus be used as a supplementary space for the main document reader, without the user having to control its display directly through their individual displays. This localised "remote control" allows for a user to, ultimately, place reading displays around their environment (as an academic may do in their office) and read from multiple sources as they compose a document. The user controls the secondary displays through a widget that then passes a message through CloudBooks to the paired reader devices. The copy of xBooks running on the reader then adjusts its display (e.g. turning a page) in response to the received message.

5 Conclusions

This paper presents CloudBooks, an architecture for reading from multiple slate PC devices. We have described systems for reading with multiple users on multiple devices in a collaborative environment, and provided support for single users allowing multiple displays to be co-ordinated through one document reader interface. The current data-model of Cloudbooks encapsulates support for location-aware and context-aware computing, allowing users to query *where* they read a text, or, conversely, what they did read in a particular time or place.

Acknowledgements. We would like to acknowledge Microsoft Research and EPSRC grant EP/F041217.

References

1. Agosti, M., Albrechtsen, H., Ferro, N., Frommholz, I., Hansen, P., Orio, N., Panizzi, E., Pejtersen, A., Thiel, U.: Dilas: a digital library annotation service. In: IWAC, pp. 91–101. CNRS (2005)
2. Amershi, S., Morris, M.: Cosearch: a system for co-located collaborative web search. In: CHI 2008, pp. 1647–1656. ACM, New York (2008)
3. Buchanan, G., Hinze, A.: A generic alerting service for digital libraries. In: JCDL 2005, pp. 131–140. ACM, New York (2005)
4. Frommholz, I., Fuhr, N.: Probabilistic, object-oriented logics for annotation-based retrieval in digital libraries. In: JCDL 2006, pp. 55–64. ACM, New York (2006)

Interconnecting DSpace and LOCKSS

Mushashu Lumpa, Ngoni Munyaradzi, and Hussein Suleman

Department of Computer Science, University of Cape Town,
Cape Town, South Africa
{mlumpa@cs.uct.ac.za,ngoni.munyaradzi@uct.ac.za,hussein@cs.uct.ac.za}

Abstract. Repository managers increasingly use toolkits such as DSpace to manage submission of and access to resources. However, DSpace does not support the highly desirable distributed replication functionality provided by LOCKSS. This paper describes an experiment to seamlessly interconnect DSpace and LOCKSS in a generalisable manner. An experimental prototype confirms that this is indeed possible, and that the interoperation can be efficient within the constraints of the systems.

Keywords: interoperability, harvesting, replication.

1 Introduction

Lots of Copies Keeps Stuff Safe (LOCKSS) [7] is a popular software toolkit to replicate digital collections. LOCKSS is noted for its voting system whereby multiple replicas of an object on distributed servers are used to ensure the authenticity of the object [4]. If a single copy of an object is corrupted, it is outvoted as the authoritative copy and updated with an authoritative copy from one of the other servers. Thus, many copies may ultimately preserve the data, as suggested by the name LOCKSS.

This replication feature is not currently considered by most institutional repository (IR) toolkits, which focus on Web-based submission and access. Thus, the administrator of an IR would need one software tool for the management of the repository (e.g., DSpace) and another for the replication function (e.g., LOCKSS).

This paper reports on an experiment to interconnect DSpace and LOCKSS, where the former is used for submission/access and the latter for replication. The feasibility of interconnecting these systems seamlessly was investigated, using an approach that is generalisable to other systems. Metadata and digital objects are then transferred between the systems as necessary using METS [6] and MODS [5] for packaging of objects and specification of metadata, with OAI-PMH [3] as the underlying transfer layer.

2 Data Harvesting

The OAI-PMH only defines the harvesting of metadata so a mechanism for access to the complete digital objects was necessary. The solution adopted was

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 493–496, 2011.

the use of a special metadata format that returns a minimal metadata record with only a URL from which the full packaged digital object can be obtained. Thus a standard OAI-PMH harvester can be used, followed by the additional step of transferring each packaged digital object individually. This solution, a minimalist version of that suggested by Van de Sompel, et al [9], is described bellow. This approach is arguably superior to METS records transferred over OAI-PMH because the packages are self-contained and can be stored and manipulated without modification of the references to objects, which would in such a solution need to initially point to the originating repository [1] [8].

The **metsPackage** metadata format is defined as containing only a single element that specifies the URL from which a METS package can be downloaded. The *metsPackage* metadataPrefix is used to request a response in the **metsPackage** metadata format.

Once the harvester has obtained this list of URLs to digital objects, it can download each package individually. These packages are in the standard METS package format and contain METS metadata in an XML file as well as all digital objects that make up the item. All files are then compressed using the ZIP algorithm.

In the experimental system that was developed, a new verb - GetPackage - was added to the OAI-PMH set to download a METS package. This was simply a development convenience - this can also be implemented using a completely different Web application. To be consistent with the OAI-PMH verbs, GetPackage takes a single parameter that is the identifier of the item whose package is being requested.

When the server creates a **metsPackage** response, it inserts GetPackage URLs into the identifier fields.

When the GetPackage verb is encountered by a server, it will dynamically create the identified METS package and return it to the harvester client. The harvester then unpackages it and ingests the metadata and digital objects into its local system.

3 Evaluation and Analysis

3.1 Experiments

The plugins were evaluated for criteria of correctness and speed. Results from the test of the speed of the LOCKSS system interface show a roughly linear increase in time to scan through the filestore as the size of the filestore increases and negligible difference in performance for shallow and deeply-nested filestores. Very small filestores appear to perform much better than larger ones, probably as a result of disk caching.A linear increase in time is expected in an unindexed data store, so this result confirms what was expected. The results for the GetPackage verb experiment show that transfer time of individual packages are reasonably consistent, with an average of 1.6 seconds per package for transfer. Variations in time are due to the differences in file sizes. The system can be deemed to be scalable with the number and sizes of files.

3.2 Post-analysis

Numerous stumbling blocks were encountered during the development that could inform the design of future repositories and systems for distributed preservation. These are discussed below.

- Read-only LOCKSS
  LOCKSS uses a read-only system partition for its operating system and all its software. This ensures that the system is stable and not prone to viruses or other attacks on the integrity of the system. Unfortunately, this also makes it impossible to add extensions to the system or update any part of the system. During the development of the LOCKSS plugin, it was necessary to copy over the plugin's files manually each time the server was restarted, as a temporary solution.
- Metadata Restrictions
  LOCKSS uses HTTP headers for its metadata while DSpace uses qualified Dublin Core. Neither system appears to allow deviation from their baselines. This poses a challenge when interconnecting the systems as metadata cannot be translated completely between HTTP headers and Dublin Core.
- Encoding structure (communities)
  LOCKSS stores structural information in the directory structure of its file-store while DSpace stores the community and collection information in its database. In both cases, there is structure in the repository but this structure is not reflected in the metadata. A solution to this is to encode the structure into the OAI-PMH set entry in a record's header. However, the structure is still not part of an offline METS package, which typically includes internal structure only. If it was possible to include structural membership information, this would still be piecemeal and empty collections would not be replicated when a repository is replicated.
- Configuration transfer
  One solution to the problem of missing structural information is to explicitly store and transfer the entire structure of a repository. This can be part of the configuration information that is necessary to restore a repository completely.

4 Related Work

Early efforts at interoperability were focused on showing the feasibility of repository-to-repository migration, for example a transfer from Greenstone to DSpace and vice versa in the Stoned project [10]. Stoned supported import and export functions for digital objects and metadata using a suite of different mechanisms.

Van de Sompel, et al [9] proposed an automated mechanism for inter-repository harvesting that is based on the OAI-PMH and the use of a complex and descriptive packaging metadata format (SCORM, MPEG-21 DIDL, METS, etc.) to represent digital objects. These metadata formats could contain

direct links to the digital objects in a live repository. Tansley [8] used a similar approach.

Finally, Duracloud promises to bridge the gap between repositories and replication by providing its own network based on cloud data storage [2] . Unlike LOCKSS, which relies on spare capacity, Duracloud will expect repository managers to explicitly allocate resources to replication. The software is open source so there is promise for experimentation with different models in the future.

5 Conclusions

This paper has described an experiment to interconnect these 2 systems to bring the benefits of both worlds closer together for repository managers.

In the process of developing an interoperability solution, the more general problem of data harvesting has been discussed. While metadata harvesting is commonplace, OAI-PMH does not address how to access the digital objects using a machine interface. This makes common services such as search engines, data mining and replication difficult. The solution adopted in this paper is not only compatible with the existing semantics of the OAI-PMH, but minimalist.

The system evaluation has demonstrated that the plugins and approach to data harvesting are both feasible and efficient. While reasonably successful, it is clear that even with the most advanced and popular tools, replication is not a simple task. Many aspects of system design can be improved to enable more effective/efficient replication, such that future attempts to interconnect similar systems will hopefully be less intensive!

References

1. Bekaert, J., Van de Sompel, H.: A Standards-based Solution for the Accurate Transfer of Digital Assets. D-Lib Magazine 11 (2005)
2. Duracloud, http://www.duraspace.org/duracloud.php
3. Lagoze, C., Van de Sompel, H.: The open archives initiative: Building a low-barrier interoperability framework. In: 1st ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 54–62. ACM, New York (2001)
4. Maniatia, P., Rosenthal, D.S.H., Roussopoulos, M., Baker, M., Guili, T., Muliadi, Y.: Preserving peer replicas by rate-limited sampled voting. ACM SIGOPS Operating Systems Review, 37–59 (2003)
5. McCallum, S.H.: An introduction to the Metadata Object Description Schema (MODS). Library Hi Tech. 22, 82–88 (2004)
6. Metadata Transmission Encoding Standard, http://www.loc.gov/standards/mets/
7. Reich, V., Rosenthal, D.S.H.: LOCKSS (Lots of copies keep stuff safe). New Review of Academic Librarianship 6, 155–161 (2000)
8. Tansley, R.: Building a Distributed, Standards-based Repository Federation: The China Digital Museum Project. D-Lib Magazine 12 (2006)
9. Van de Sompel, H., Nelson, M.L., Lagoze, C., Warner, S.: Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine 10 (2004)
10. Witten, I.H., Bainbridge, D., Tansley, R., Huang, C.Y., Don, K.: A bridge between greenstone and DSpace. D-Lib Magazine 11 (2005)

Encoding Diachrony: Digital Editions of Serbian 18th-Century Texts

Toma Tasovac[1] and Natalia Ermolaev[2]

[1] Center for Digital Humanities, Belgrade, Serbia
ttasovac@humanistika.org
[2] Program in Library and Information Science
School of Communication and Information
Rutgers, The State University of New Jersey
New Brunswick, NJ, USA
n.ermolaev@rutgers.edu

Abstract. Texts in the "Digital Library of Serbian Cultural Heritage of the 18th Century" are encoded as a word-aligned corpus of TEI XML documents in two versions: one using traditional 18th-century orthography, including the graphemes which have since disappeared from Serbian, and one using modernized and standardized Serbian spelling rules that increase the legibility and searchability of these texts for modern users. The corpus also contains linguistic and semantic annotations that add modern phonetic, morphological, lexical and conceptual equivalents to the largely archaic vocabulary. By applying basic techniques of cross-lingual information retrieval to a historical dimension of one language, and making provisions for multiple indexing and annotations, our project exposes a notoriously difficult chapter in the development of the Serbian language to a wider audience, without sacrificing the edition's scholarly potential.

Keywords: Digital humanities, digital editions, digital libraries, cultural heritage, Serbian language, language change, Dositej Obradović, Text Encoding Initiative (TEI).

1 Serbian 18th Century: Linguistic and Technological Challenge

While many European languages underwent relatively uninterrupted linguistic and cultural evolutions, the modern Serbian literary language, as codified by Vuk Stefanović Karadžić in the 19th century, had its roots in the vernacular rather than the literary standard(s) of the previous epoch [1]. 18th-century Serbian language itself was largely influenced by three linguistic forces: 1) the ecclesiastical Serbian recension of Church Slavonic (Serbo-Slavonic); 2) the Russified recension of Church Slavonic known as Ruso-Slavonic; and 3) a fashionable hybrid of Ruso-Slavonic and vernacular Serbian known as Slaveno-Serbian [2]. The linguistic complexity of these texts, therefore, poses a formidable challenge to the modern reader, who is unaccustomed to archaic vocabulary and morphology of the pre-reform Serbian.

Our initial digitization focus has been on the writings of Dositej Obradović (1740-1811), widely considered the chief representative of 18th-century Serbian and South Slavic Age of Enlightenment [3].

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 497–500, 2011.

1.1 The Problem with Print Editions

Today, the works of 18th-century Serbian authors generally, and Dositej Obradović in particular, are published exclusively in their modern-day transcription: the original orthography is standardized and modernized, while lexical and morphological peculiarities remain untouched [4]. These print editions usually include a glossary of lesser-known words listed in the appendix, but the annotated words are not marked in the text itself, which makes the process of consulting the glossary highly impractical, time-consuming and often frustrating.

2 Making User-Centered Digital Editions of 18th-Century Texts

To be truly useful for any kind of textual study, a digital library must provide a range of options – e.g. full-content searchability, concordances and indexes, metadata, hyperlinks, and critical markup [5]. When dealing with 18th-century Serbian texts, however, we must ask if traditional indexing techniques are suitable for works written in non-standardized, pre-reformed Serbian orthography. Even if the archaic vocabulary of the transcribed originals is manually lemmatized, only a handful of scholars would profit from their full-text search functionality. With these factors in mind, we have chosen an editing strategy that allows us to maximize both the scholarly and the non-academic appeal of the obscure yet rich linguistic fabric of 18th-century Serbian.

While it is important to preserve (i.e. not resolve) the non-standard textual features (word segmentation, abbreviations, variants of letters and punctuation marks) in critical editions [6], linguistic puritanism will not suffice if we aim to expose cultural heritage objects to a broader audience. We are therefore creating hybrid digital editions by: a) encoding texts both in their original orthographic form and their modern-day, standardized equivalent; b) annotating phonetic, lexical and conceptual equivalents of words which are no longer used in contemporary Serbian; and c) developing a web application and browser interface for reading/searching the texts, providing several levels of indexing to enable multiple access points. We are applying basic techniques of cross-lingual information retrieval (CLIR) to a historical dimension of one language.

2.1 Encoding

The core of our digital edition of Dositej Obradović is a parallel corpus of his texts in their *original orthography* [OO], and transcriptions of those texts using standardized Serbian *modern orthography* [MO]. Texts are edited in XML according to the Guidelines of the Text Encoding Initiative (TEI) [7]. The corpus is word-aligned and heavily indexed so that each word, sentence, paragraph and division is assigned a unique ID as an anchor for establishing corresponding links between the two transcriptions.

2.2 Lexical and Semantic Annotation

We have established the following guidelines for lexical and semantic annotation of 18th-century texts:

1. If the difference between a Slavonicism and its modern equivalent is purely orthographic, i.e. if MO is also equal to the modern Serbian word, (e.g. богъ vs. бог, сладка vs. слатка etc.), no special annotation is provided.
2. If a Slavonicism (e.g. любовъ [OO]; љубов [MO]) is phonetically distinct from its modern Serbian counterpart (љубав), but the two are otherwise semantically equivalent, the Slavonicism is annotated with its modern Serbian *phonetic* equivalent.
3. If a Slavnonicism (e.g. утѣшенїе [OO]; утешење [MO]) and its corresponding modern-day Serbian variant (утеха) have the same root, but different morphological realizations, the Slavnonicism is annotated with its modern-day *morphological* equivalent.
4. If a Slavnonicism (e.g. благодѣтель [OO], благодетељ [MO]) is no longer used in contemporary language, but there is a corresponding lexical equivalent (i.e. synonym or near-synonym, e.g доброчинитељ), the Slavnonicism is annotated and linked to its modern Serbian *lexical* equivalent.
5. Finally, if a Slavnonicism is not lexicalized in the modern language (e.g. любезница [OO]; љубезница [MO]), a modern Serbian *conceptual* equivalent (драга особа женског пола) is assigned to the archaic form and marked as such.

2.3 User Interaction with the Digital Library

Our encoding will allow the reader to search the contents of the digital library using not only the modern Serbian orthographic forms [MO] of the original texts, but also the phonetic, lexical and parsed conceptual equivalents of the words appearing in the original. For example, a search for the modern-Serbian lexeme захвалност (gratitude) would also return and indicate examples containing the Slavonicism припознанство, even though the modern-day lexeme only occurs in annotations and is not actually present in either of the two transcriptions of the text.

Our system of diachronic annotation is important for several reasons:

1. With a language of low-standardization, such as 18th-century Serbian, searching the text via modern-day equivalents can retrieve a larger pool of data than a search based on original orthographic forms.
2. 18th-century Serbian employs graphemes no longer present in modern Serbian (for instance я, ю, ї, ъ, щ etc.) and not readily available in Serbian Cyrillic keyboard layouts.
3. 18th-century Serbian vocabulary is largely unknown to an average native speaker and student of the Serbian language; searching via lexical equivalents across different works and authors may uncover previously unnoticed thematic similarities and correspondences.

3 Conclusions and Further Work

Even though additional levels of markup – such as part of speech or syntactic frames – could be manually applied to our digital editions, we are currently focusing on the above guidelines in order to gather a pool of texts that could also be used for experimenting with NLP techniques such as training a tagger for semi-automatic

POS-tagging. Another possibility is the automatic creation of bilingual glossaries based on our annotations, and the cross-referencing of 18th-century sources with the synonym-rich Transpoetika Dictionary of the Serbian Language [8] – which would help us create even more sophisticated search mechanisms for 18th-century texts. Providing both original and standardized transcriptions of heavily annotated 18th-century text will open new areas of research and application of digital technologies in the study of this important period in Serbian history.

References

1. Ivić, M.: O jeziku Vukovom i vukovskom. Novi Sad: Knjiž. zajednica Novog Sada (1990) (in Serbian)
2. Albin: The Creation of the Slaveno-Serbski Literary Language. The Slavonic and East European Review 48, 483–491 (1970)
3. Jovanović-Gorup, R.: Dositej Obradović and Serbian Cultural Rebirth. Serbian Studies 6, 45–80 (1991)
4. Stefanović, M. (ed.): Sabrana dela Dositeja Obradovića , pp. 1-6. Zadužbina "Dositej Obradović", Beograd (2008)
5. Tasovac, T.: Why every picture is not worth a thousand words: digital libraries from a textual perspective. In: Vraneš, A., Marković, L.J., Vulović, K. (eds.) Electronic Library, pp. 721–732. Filološki fakultet, Beograd (2009) (in Serbian)
6. Pusch, C., Kabatek, J., Raible, W. (eds.): Romance corpus linguistics II: corpora and diachronic linguistics. Gunter Narr Verlag, Tübingen (2005)
7. TEI Consortium, (eds.): Guidelines for Electronic Text Encoding and Interchange, `http://www.tei-c.org/Guidelines/P5/`
8. Tasovac, T.: More or Less Than a Dictionary? Wordnet as a Model for Serbian L2 Lexicography. Infotheca: Journal of Informatics and Librarinaship 10, 13a–22a (2009)

Cross-Border Extended Collective Licensing: A Solution to Online Dissemination of Europe's Cultural Heritage?

Johan Axhamn*

johan.axhamn@juridicum.su.se

An issue which recently has gained increased attention from legislators is how to stimulate the digitization and online availability of the collections held by libraries, museums and other cultural institutions – sometimes referred to as our "common heritage" – and at the same time give full respect to established copyright norms. At European level, this attention is evident in the Digital Libraries Initiative, the Communication from the European Commission on Copyright in the Knowledge Economy, the Commission's Digital Agenda for Europe and its recent Communication on a Single Market for Intellectual Property Rights. Inherent in these policy documents is the recognition that the new information technologies have created vast opportunities to make the common heritage of Europe more accessible for users online. It is also a shared belief that such access – if coherent with basic copyright principles – will be for the mutual benefit of users, right holders and the society at large. In line with this the Commission has supported the creation and development of a common access point for Europe's cultural heritage, *Europeana*.

However, several issues from a copyright perspective have to be solved before undertaking mass-digitization and online dissemination of the collections held by these institutions. One of them is how to make the said digitization and online dissemination lawful from a copyright perspective. To the extent that an item in a cultural institution's collection is (still) protected by copyright, those acts fall under the author's exclusive right to authorize and prohibit use of his or her work. As the administrative ("transaction") costs of finding and negotiating an individual license with every right holder would rise to astronomical levels, there is an obvious risk that major parts of the collections will not be digitized and disseminated online. For this reason, the most practical solution would probably lie in the area of collective rights management.

A way forward is the extended collective license (ECL) model as established and developed in the Nordic countries. The essential component of the ECL model is that it extends a freely negotiated agreement between a Collective Management Organization (CMO) and a user so that it binds also non-members' rights, sometimes referred to as "outsiders' rights". The legal implication of this extension effect is that the agreement not only gives the user the right to use outsider's rights without any risk of civil remedies but that that it also provides full limitation against criminal sanctions. To safeguard the outsiders' interests, the legally supported extended effect only occurs provided that certain conditions have been met. These conditions are, mainly, outsiders' possibility to opt out, equal treatment vis-à-vis members of the organization and receipt of remuneration. There are also conditions related to the representative-

* Faculty of Law, Stockholm University.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 501–504, 2011.

ness and supervision of the eligible CMOs. The ECL model has been under consideration by the Commission as a possible solution to stimulate the digitization and online availability of the collections held by cultural institutions.

An additional challenge is to make the collections available cross-border, i.e. also to other countries (territories) than the one where the cultural institution is located. It is inherent in the policy documents of the European commission and also the establishment of *Europeana* that there is a clear political aim to stimulate such cross-border dissemination. According to prevalent copyright rules, rights for dissemination online have to be cleared in every country where the content can be accessed. Applied to cultural institutions this means that they would have to get a license from CMOs in every EU member state. This would of course lead to substantial administrative costs for the institutions. However, so far no solution have been brought forward which takes into account and could be acceptable by both cultural institutions and right holders.

Two viable cross-border solutions are a country of transmission principle or a solution based on voluntary measures by the national CMOs. A country of transmission principle holds that cultural institutions should only be obliged to obtain a license in the country where the institution initiated the online dissemination. This solution would require legislative intervention at EU level. The other solution essentially means that national CMOs would give each other a mandate to issue multi-territory licenses.

At first glance, an ECL provision combined with either of the cross-border solutions outlined above may be regarded as favoring the cultural institutions' interests, as it gives them the privilege of both an ECL provision and a simplified measure for cross-border rights clearance. However, the scope of an ECL provision for the benefit of these institutions would primarily be to make available content that is not of a contemporary commercial nature. Hence, the model would aim at establishing a mechanism which would create a supply of cultural heritage content. It is in the interest of the society as a whole that also this content is made available online.

Against this background, the panel will discuss pros and cons of a cross-border ECL model in relation to the digitization and online cross-border availability of the collections held by national cultural institutions. The panelists are representatives of different interests (stakeholders).

Moderator

Patrick Peiffer is project manager for the Services électroniques at the National Library of Luxembourg, managing licences for the national consortium and specialising in copyright issues, currently for retro-digitisation of newspapers and digital legal deposit. He is member of the EU Member States Expert Group on Digitisation and within the EuropeanaConnect project, task leader for the Europeana Licensing Framework.

Panelists

Annemarie Beunen is the copyright lawyer of the National Library of the Netherlands, and an assistant professor at Leiden University (department of eLaw). Here, she lec-

tures and publishes on copyright issues relating to digitised cultural heritage. Annemarie read Dutch Law (specializing in copyright) and History of Art at Nijmegen University in the Netherlands. In 2007 in Leiden, she finished her PhD thesis on the sui generis protection for database producers under the European Database Directive. In 1999-2000 she edited the first Dutch legal guidebook for museums. She also held former positions at the Dutch Council for Culture, and the Dutch Council of State.

Johan Axhamn is a PhD candidate in intellectual property law at the Faculty of Law, Stockholm University, where he conducts research and is a teacher in intellectual property law. His PhD project deals with the EU database directive, especially the sui generis database right and its interfaces with competition law and fundamental rights. Between November 2010 and May 2011 Johan conducted research at the Institute for Information Law (IViR) at the University of Amsterdam on the Nordic ECL model in relation to the Europeana project. This research has resulted in a EuropeanaConnect ECL report on practical solutions for cross-border access, available in Q3 2011. Johan is a former special adviser to the Ministry of Justice in Sweden, with special focus on copyright and enforcement issues and is currently Sweden's expert and delegate to the WIPO Intergovernmental Committee on Intellectual Property and Genetic Resources, Traditional Knowledge and Folklore.

Silke von Lewinski is tenured at the Max Planck Institute for Intellectual Property, Munich and specialises in international and European copyright law. She is also an adjunct professor at Franklin Pierce Center for IP at the University of New Hampshire Law School, Concord, N.H., USA. Dr. von Lewinski frequently has been an expert consulting the European Commission, e.g. on the EC Rental Rights Directive, and regarding the WIPO Diplomatic Conference 1996. At the WIPO Diplomatic Conference 2000 on Audiovisual Performances, she was a delegate for Germany. She has been the chief legal expert consulting the governments of Eastern and Central European and former Soviet countries on their copyright legislation in the framework of the initial EC's TA programs from 1995 and works under subsequent programs, including in Asia. Numerous publications and lectures worldwide have focussed on copyright law, primarily international and European. Main publications: "The WIPO Treaties 1996" (2002, with Reinbothe), "International Copyright Law and Policy" (OUP, 2008), "European Copyright Law" (OUP, 2010, with MM Walter et al.); edited: "Indigenous Heritage and Intellectual Property: Genetic Resources, Traditional Knowledge and Folklore" (2nd edn. 2008); "Copyright throughout the World" (West, from 2008).

Mette Møller is the Secretary General of the Norwegian Authors' Union. She graduated from the University of Oslo in 1991 and has worked within the copyright field ever since. Møller participated in the working group for the test case to bokhylla.no ("Bookshelf"), a project aiming at making all Norwegian books from the 1790s, 1890s and 1990s available on the Internet. To the extent that books are still under copyright, they are encompassed by an extended collective licensing agreement between the Norwegian collective management organization Kopinor and the National Library of Norway. Møller has followed bokhylla.no closely since it was launched in 2009. Møller also has experience from working at the Norwegian Ministry of Cultural Affairs with copyright policy issues in both a Nordic and European framework, con-

tributing to a revision of the Norwegian Copyright Act in 1995. She has also worked as a Legal Manager at Norwaco (a collective management organization clearing rights and collecting remuneration for retransmission of broadcasting signals with legal basis in an extended collective license agreement). Between 1996 and 2006 she worked as an IPR business lawyer. Between 2005 and 2009 she was a Board member of the Norwegian Film Fund.

An Investigation of ebook Lending in UK Public Libraries

Christopher Gibson

Department of Computer and Information Science, University of Strathclyde, Glasgow
cg@cis.strath.ac.uk

Abstract. This research aims to investigate ebook lending, management, and procurement in UK public libraries. A mixed method approach will be utilised to gain an understanding of how ebook lending is currently being achieved and to determine its affect on traditional library services. This research also proposes to run ebook reader lending trials from selected public libraries.

Keywords: ebooks, ebook readers, public libraries.

1 Introduction

The ebook revolution is coming; driven by innovative business models, supported by impressive technological developments, and fuelled by apparently endless content. Despite difficulties in obtaining reliable data on ebook sale figures [1], evidence from the International Digital Publishing Forum suggests that sales of ebooks in America are growing rapidly, from 6 million dollars in 2002 to 55 million dollars in 2009 [2]. Public libraries are receiving active encouragement from government to provide ebook services [3] but face lukewarm responses by library users [4] and a volatile sector in which Google, Amazon, Apple, and a host of other organisations fight for a share of the marketplace. The provision of ebook services in the academic library sector is relatively widespread - over 60% of students utilised them in their studies in 2008 [5] - but while ebook services have been established in US public libraries for a number of years [6], the available data suggests that the vast majority of UK public libraries are yet to offer ebook services [4]. The British Library recently estimated that by 2020 the vast majority, 95%, of new works published will only be published in electronic form [7].

Despite a significant body of research on the design of ebooks [8], and their use in an academic context [9, 10] there is a dearth of research concerning ebook lending services in public libraries. The objectives of this study are to investigate ebook service provision, both within public libraries and other library sectors, in order to develop an understanding of the myriad challenges involved and to provide guidance on how collection development policy must evolve to meet new demands. The potential benefits of ebook service in public libraries would be an actualisation of the 24/7 virtual library for all users as well as being of significant interest to specific user communities, such as the visually impaired, reading groups [11], and the young. This research will be of interest to library practitioners wishing to introduce or optimise ebook lending services.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 505–510, 2011.

2 Research Questions

The research intends to examine current and future practice relating to ebook lending provision in the UK in order to develop a best practice model that will be useful to public library practitioners wishing to develop or enhance such services. The research questions are as follows:

Q1. How have public libraries addressed ebook service provision in the UK?
Q2. What challenges and opportunities exist in incorporating ebook lending into other reader services?
Q3. Is it feasible to lend ebook reading devices from public libraries?
Q4. How can the effectiveness of ebook lending services be measured?
Q5. How do library users view the provision of ebook lending services?
Q6. How can effective ebook lending services be developed?

3 Related Work

Whilst ebooks are now commonly found in academic libraries [5] they are yet to fully penetrate public libraries in the UK [4]. Ebooks in public libraries have attracted less research interest than ebooks in academic libraries; the majority of work in the UK has been undertaken by James Dearnley and Cliff McKnight of the Department of Information Science at Loughborough University. Public libraries in the UK lag behind the US in offering ebook lending services [6], in the US public libraries first offered ebook lending through the aggregator netLibrary in 1998 (Genco, 2009). The first public library authorities to lend ebooks in the UK were Richmond [12] and Blackburn and Darwen [13]. Whilst research into ebooks in academic libraries has recently been characterised by large scale surveys [5, 14] the equivalent work in the public library arena has been on a smaller scale. Recently more attention has been paid to ebooks in public libraries and Highwire Press of Stanford University [15] released a report that included a large number of public libraries in the sample. As public libraries increasingly begin to adopt ebooks and offer digital lending it is likely that interest from the scholarly community will follow. Currently the majority of public libraries in the UK do not offer ebook lending services but the number is increasing rapidly. It seems that the aggregator Overdrive has a commanding market share amongst public libraries in the UK and the US [12] but this is liable to change in a highly volatile market place [1]. Public libraries face a number of challenges in the following areas when offering ebook lending: selection, acquisition, cataloguing, access, preservation, and management.

Dearnley and McKnight [16] conducted two pilot studies that investigated the potential usage of Rocket eBook ebook readers in a public library context. The first of these studies was carried out at Loughborough University and the second at Market Harborough public library. It resulted in a trial period of lending ebook readers in Market Harborough public library. The study was ostensibly an evaluation of ebook lending in a public library context but was more accurately a usability evaluation of the Rocket eBook ebook reader and Glassbook ebook reader software. The conclusions drawn at the end of the study were that ebook readers would have to be significantly improved before ebooks achieved mainstream acceptance.

Manyard and McKnight [17] investigated the provision of ebooks for children in public libraries. They found a positive attitude towards ebooks as a method of enhancing the children's services that the library offered. A common problem identified by the librarians in the survey was that lending ebooks was limited by the possibility of children not having access to hardware that would allow them to use the ebook.

Dearnley and McKnight [13] published a further paper in 2003 that again explored ebook use in a public library context. An overview of ebook lending in the US was provided and the conclusion was reached that the situation in the UK was less well developed owing to the limited availability of ebook readers, lack of UK specific ebooks, and the lack of commercial ebook vendors. It was found that lending the ebook readers was problematical for staff and the Rocket eBook was not a suitable device. However the paper concluded that a new generation of ebook reading devices could offer benefits to public libraries.

Vidana [18] provided insight into the process that the library authority would have to undertake in order to prepare for lending ebooks: choosing ebooks providers and formats, selecting titles, resolving technical issues, training staff, developing evaluation processes, launching and publicising the service, and monitoring and assessing the progress. Vidana also mentioned the possibility of using ebooks to form consortia which is currently a strong trend in the US [19].

Garrod [6] gave an overview of the ebook market place, which she conceded was a complex topic. Garrod notes the importance of marketing the service, which has been a common theme in the promotion of ebooks in academic libraries [20] as well as suggesting consortia based solution to collection management.

Dearnley, McKnight, and Morris [21] report on user and staff reactions to a personal digital assistant (PDA) based ebook collection. The conclusion that ebook were to be viewed as supplementary to physical collections was also aired. Again it was found that the hardware used in the study was not sufficiently usable as to be practical to be lent from the public library.

Dearnley, McKnight, and Morris [4] present the results of an online questionnaire survey that collected data on the ebook collection held at Essex County Libraries. The collection was provided by Overdrive and Ebrary. The conclusions reached was that marketing of ebook collections was paramount in their adoption by users and that certain genres, especially science fiction, were more popular than others. Again the study concludes optimistically with the promise of further advances in ebook reader technology driving the ebook market forward.

Landoni and Hanlon [11] used two reading groups in public libraries in Glasgow to explore the utility of fiction ebooks. Despite the negative reactions for the participants Landoni concludes that there 'is undoubtedly a role for e-books in the public library service'.

The scale of these various studies differs but is generally fairly low; the largest of the studies surveyed 58 people [4]. Despite the low numbers involved the conclusions from the studies are encouraging and this researcher believes the time is right to revisit the issue of ebook lending provision in public libraries, particularly now that many libraries are beginning to loan ebooks. The methodologies of the studies described will be used to inform the researcher's own, particularly in regard to the nuts and bolts planning of the ebook lending trial described below.

4 Anticipated Methodology

The anticipated methodology consists of freedom of information (FOI) requests to gather quantitative data on the state of ebook lending provision throughout the UK. These requests would be used to gather data on a range of issues and would, amongst other valuable insights, give an idea of the scope and penetration of ebook lending provision, the leading third party providers of these services to libraries, and the current level of use of these services. Other interesting information would be the popularity of individual titles or genres to get a better idea of how the library user borrows ebooks. It is important to consider how a representative sample of library authorities can be gathered as well as the practicalities of gathering the information, such as who the FOI request should be made to and whether to do this in paper or by email. The risk associated with this data gathering method is that the library authorities contacted with the FOI might resist releasing data either by claiming that the information is commercially sensitive or that the request would take too much time to complete. Careful formulation of the FOI request and the development of strategies to deal with obfuscation would combat this anticipated risk.

A case study approach to gathering data would be rewarding in developing links to particular library authorities and may prove crucial in carrying out the data gathering exercises detailed below. The possibility of forming a strategic partnership with a library authority will be given consideration as it may facilitate data gathering.

It is also envisioned that interviews with library practitioners will play an important role in gathering qualitative data on the current and future practice when lending digital material through a public library. Interviews with other interested parties, i.e. ebook publishers and aggregators, may also prove to be useful in answering the research questions. Another way to gather this data may be through the use of questionnaires. The risk associated with this method of gathering data is that it may not be possible to arrange interviews around interviewees' busy schedules. Questionnaires may also be ignored. In anticipation of this plenty of time is being allocated to field work in order to give the best chance of arranging mutually agreeable interview slots.

It is possible that the methodology will include a trial that involves lending ebook reader devices in conjunction with a library authority. The researcher's department is in possession of various ebook readers and it is hoped that these devices could be utilised as an incentive for a library authority to participate in the trial. This would simulate ebook lending from a public library and be based on work undertaken by Landoni and Hanlon [11] but would be significantly larger. This would lead to information being gathered on the practical realities of lending ebook readers through pubic libraries, as well as the experience of users of ebook lending services, and would likely included a cost-benefit analysis. A further benefit would be to put the researcher in touch with public library users who utilise ebook lending services. There are a number of potential risks to this method of data gathering; the first major obstacle is getting a library authority to agree to take part. It will also be difficult to ensure representativeness in a trial of this sort as the participants may not cover the full spectrum of public library users, although Dearnley and McKnight [4] achieved a surprisingly broad demographical spread when conducting their research. There may also be financial implications in lending ebook readers from a public library in terms of having them insured for such a use.

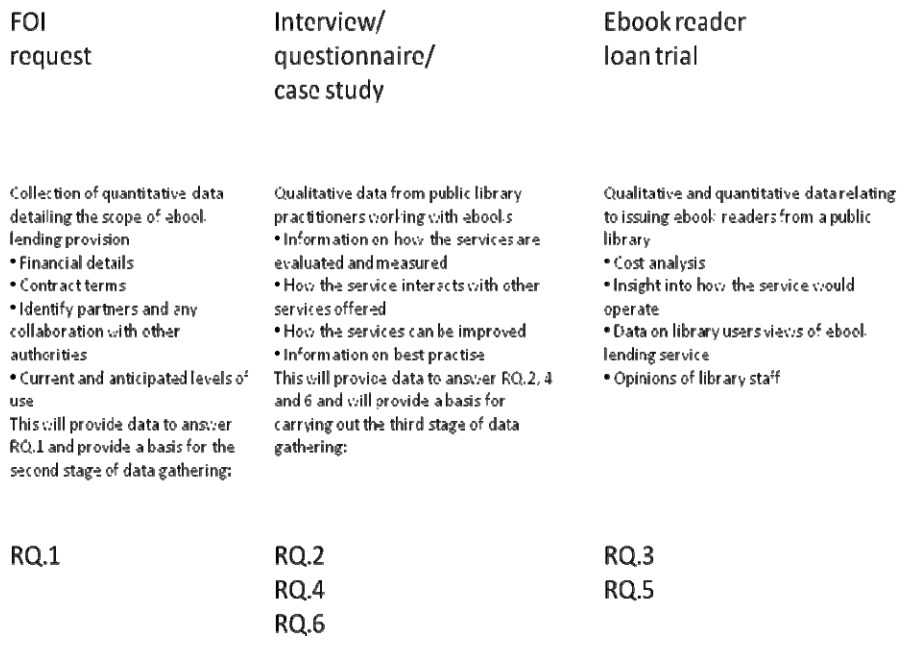


Fig. 1. Anticipated methodology

5 Anticipated Outcomes

Anticipated outcomes of this research include gaining an understanding of current practice in the UK. By mapping the current state of play it is thought that a best practice model could be developed aiding public library practitioners wishing to introduce or optimise ebook lending from their library. This research has the potential of offering a great deal of knowledge exchange, both with librarians and with private enterprise seeking to provide third-party support to libraries offering ebook lending. It is anticipated that library users will benefit from the opportunity to air their opinions about ebooks.

6 Issues for Discussion

The researcher is currently working on firming up the research methodology and would appreciate comments relating to methodological approach. Although the methodology is currently at a very early stage it is certain that by the time of the doctoral consortium it will have advanced considerably and some initial results from the fieldwork will be available. Initial results suggest that public libraries are considering building their own ebook libraries from diverse sources; it would be interesting to discuss methods of achieving with this with the panel. Other possible issues for discussion include the various business models related to lending ebooks, possible disintermediation of public libraries by third parties, the rise of consortia, and the recent Google books ruling.

References

1. Vasileiou, M., Hartley, R., Rowley, J.: An overview of the e-book marketplace. Online Information Review 33(1), 173–192 (2009)
2. US Trade Wholesale Electronic Book Sales.: International Digital Publishing Forum (2009), http://www.openebook.org/doc_library/industrystats.htm
3. Jefferies, S.: The Battle of Britain's Libraries. The Observer (March 7, 2010)
4. Dearnley, J., McKnight, C., Morris, A.: Making e-books available through public libraries: some user reactions. Journal of Librarianship and Information Science 40(1), 31–43 (2008)
5. Nicholas, D., et al.: UK Scholarly e-book usage: a landmark survey. Aslib Proceedings: New Information Perspectives. 60 (4), 331-334 (2008)
6. Garrod, P.: E-books in UK libraries: Where are we now? Ariadne 37 (2003), http://www.ariadne.ac.uk/issue37/garrod/
7. British Library: 2020 Vision (2010), http://www.bl.uk/2020vision
8. Landoni, M., Gibb, F.: The role of visual rhetoric in the design and production of electronic books: the visual book. The Electronic Library 18(3), 190–201 (2000)
9. Wilson, R.: Ebook Readers in Higher Education. Educational Technology and Society 6(4), 8–17 (2003)
10. Jamali, H., Nicholas, D., Rowlands, I.: Scholarly e-books: the views of 16,000 academics. Results of the JIC national E-Book Observatory. Aslib Proceedings: New Information Perspectives. 61(1), 33–47 (2009)
11. Landoni, M., Hanlon, G.: E-book reading groups: interacting with e-books in public libraries. The Electronic Library 25(5), 599–612 (2007)
12. Genco, B.: It's been Geometric!! Documenting the Growth and Acceptance of eBooks in America's Urban Public Libraries. In: IFLA, Milan (2009)
13. Dearnley, J., McKnight, C.: Electronic book use in a public library. Journal of Librarianship and Information Science 35(4), 235–242 (2003)
14. Abdullah, N., Gibb, F.: Students' attitudes towards e-books in a Scottish higher education institute: part 1. Library Review 57(8), 593–605 (2008)
15. Highwire: 2009 Librarian eBook Survey (2010), http://highwire.stanford.edu/PR/HighWireEBookSurvey2010.pdf
16. Dearnley, J., McKnight, C.: The revolution starts next week: the finding of two studies considering electronic books. Information Services and Use 21, 65–78 (2001)
17. Manyard, S., McKnight, C.: Electronic books for children in UK public libraries. The Electronic Librarian 19(6), 405–423 (2001)
18. Vidana, M.: Why e-books? Library and Information Update (2003)
19. COSLA: eBook Feasibility Study for Public Libraries (2010), http://www.cosla.org/documents/COSLA2270_Report_Final1.pdf
20. Taylor, A.: E-books from MyLibrary at the University of Worcester: a case study. Program: Electronic Library and Information Systems 41(3), 217–226 (2007)
21. Dearnley, J., McKnight, C., Morris, A.: Electronic book usage in public libraries: a study of user and staff reactions to a PDA-based collection. Journal of Librarianship and Information Science 36(4), 175–182 (2004)

Leveraging EAD in a Semantic Web Environment to Enhance the Discovery Experience for the User in Digital Archives

Steffen Hennicke

Humboldt-Universität zu Berlin,
Berlin School of Library and Information Science,
Dorotheenstr. 26, 10117 Berlin, Germany
steffen.hennicke@ibi.hu-berlin.de

Abstract. The proposed study investigates the information needs and information-seeking behavior of archival users. For this purpose the ARGUS information system of the German Bundesarchiv and related reference questions are analyzed in a case study in order to model patterns of questions and search behavior in an ontology. This knowledge graph represents the knowledge archival users expect from archival finding aids. It is being compared with the knowledge graph of archival finding aids encoded with the *Encoded Archival Description* (EAD) standard in order to identify semantic gaps. The aim is to find out if information modeled in EAD matches the archival user's expectations and to formulate a model and methodology which can be applied and validate in similar cases of digital archives in order to improve and facilitate access to archival information systems.

Keywords: information need, information-seeking behavior, archival user, archive, user study, EAD, semantic web, finding aid, holding guide, archival reference question.

1 Research Interest

The main means of access to primary sources in an archive are finding aids supported by the expertise of archivists. Archival finding aids typically describe a fonds and the records it contains according to the principle of provenance.[1] A record does not denote a single object like in a library catalogue but the - possibly many and diverse - contents of a "box" and its context within the particular fonds.

Archival finding aids are only useful to a researcher if the user's cognitive representation of an archive converges with the archivist's cognitive representation of the fonds [5]. This refers to one of the essential issues of the archive's relation

[1] This fundamental archival principle emphasizes custodial history and takes into account the institutional origin and original order of items and their context in a collection now deposited in the archive.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 511–514, 2011.

to its users which is the convergence of expectations and needs of archival users with the traditional archival documentation practice and customary archival access tools. Studies show prevailing discrepancy in how archival information should be structured, ordered, and presented [1].

Therefore it is essential to gain deeper knowledge of the information needs of the archival user by studying reference questions with which users approach archives [3]. The analysis of the information-seeking behavior of users - but also of archivists - in pursue of answers to those questions is a necessary second step in order to thoroughly comprehend how information needs in the form of questions are applied to archival finding aids [4]. Together, they constitute the knowledge structure users expect to find in an archival finding aid. The *Encoded Archival Description*[2] (EAD) standard is the latest and most promising effort to bring standardization to archival finding aids and their online publication. However, it is an open question how well its knowledge structure matches the expectations of archival users.

2 Research Questions and Approach

The main research question is: Does information encoded in EAD have semantic gaps that impede matching users' information needs to archival records?

The graph-based approach of the *Semantic Web*[3] allows to represent user expectations towards an archival information system in an ontology and compare this knowledge structure to the one of EAD in order to identify semantic gaps which would impede satisifiying users' information needs. At the same time, such an ontology would constitute a richer context layer around EAD-encoded archival finding aids providing records with "anchors" for further contextualization and thereby a means to close semantic gaps.

The following secondary (implicit) research questions need to be tackled in order to answer the main research question: (1) What information needs do archival users articulate towards archives in the form of reference questions? (2) How do archival users and archivists seek for answers in digital archival finding aids? (3) How do archival users apply their information needs to an archival information system? (4) How can Semantic Web technology enhance and facilitate discovery and retrieval in archival information systems? (5) Is the hierarchical archival description compatible with a graph-based representation?

3 Research Plan and Methodology

3.1 Research Data

Reference questions are provided by the Bundesarchiv who provide their archival information system ARGUS[4] which is the first of its kind in Germany to use

[2] http://www.loc.gov/ead/ [2011/06/04]
[3] http://www.w3.org/standards/semanticweb/ [2011/03/25]
[4] http://startext.net-build.de:8080/barch/MidosaSEARCH/search.htm [2011/03/20]

EAD. The system integrates archival finding aids into a common search engine and search interface and contains 1.600+ finding aids with 1,5+ million records. The records include a broad range of administrative legacy from German history since 1495. About 3000 files from different records have been digitized and encoded in METS[5].

3.2 Users Expectations towards Archival Information Systems

This dissertation first gains a deepened understanding of information needs and information-seeking behavior of archival users by looking at the process of stating and satisfying archival information need. First, a set of reference questions will be analyzed. This analysis adapts the approach of Duff and Johnson [3]. The aim is to find out about the type and structure of archival questions and the terms used and their semantic relations. The second part will validated and complemented the results of the first part by conducting an experiment which focuses on the information-seeking behavior within the specific archival information system ARGUS. Users will be asked to perform a range of typical tasks and seek answers to previously identified typical questions. Data collection methods will include a transaction log analysis. The aim is to discover trends in the information-seeking behavior and specifically to find out about term patterns in query formulation and navigational patterns within the archival aid structure and the search results. The first and second parts of the dissertation provide the data for conducting the third part.

3.3 Identifying Semantic Gaps

The leading research question in the third part is if the information that can be encoded in EAD, i.e. is encapsulated in its very structure, is sufficient to answer to the user questions and to meet their expectations. First, EAD will be converted to a graph-based representation. Secondly, the knowledge users expect from an archival information system, will be represented in an ontology. At the moment, the most promising candidate for both tasks is CIDOC-CRM[6] which is a high-level ontology to enable information integration for cultural heritage data and their correlation with library and archive information [2]. By comparing both representations we can identify missing key elements, i.e. semantic gaps, in EAD and shortcomings in its traditional hierarchical structure as it is. This comparison will be conducted as a case study using a limited selection of finding aids from the ARGUS information system.

A concrete example for a semantic gap would be a search for the name of a municipality which might be present in the full text of an EAD element but in a different or outdated spelling or language than the user typed in. In such a case the municipal name and corresponding records would not be found. Also the territory or the national ownership of the municipality itself could have changed

[5] http://www.loc.gov/standards/mets/ [2011/06/04]
[6] http://www.cidoc-crm.org/ [2011/03/20]

during the course of history. Named entity recognition and semantic enrichment and alignment to a controlled vocabulary could help to close such a semantic gap. Another case would be when the user found a record, for example about a person, and now the user is interested in records which are related to the person mentioned in the initial record, for example by occupation, profession, family, birth/death date/place etc. The user could also be interested in similar relations to the record itself, for example by creator, creation date, provenance, custody etc. Such connections might be implicitly available in the EAD knowledge structure located at different levels of the descriptive hierarchy but without being explicitly expressed they cannot be exploited. Such missing connections are semantic gaps which can be closed by applying a knowledge structure on top of EAD. Full text search fails to discover such connections and to place the information into context found in the descriptive hierarchy.

4 Summary: Aims and Value of the Study

The dissertation contributes to the deficient understanding of the archival user needs. Furthermore, it represents user expectations in an ontology and matches this knowledge structure to the EAD knowledge structure in order to identify semantic gaps which impede matching users' information needs to archival records. This comparison is captured in a methodology which can be applied to and/or validated in similar contexts of other digital archives. It is important to note that this research does not try to define a new archival description standard or to revise an existing one but to formulate a mediator, i.e. a boundary object, as an additional layer on-top of EAD. The outcome of this study, therefore, will be on the one hand the suggestion of an ontology which is meant to facilitate research in terms of searching, finding, and discovering information in archival information systems and on the other hand a methodology build around modeling user expectations into such an ontology which allows to identify semantic gaps in EAD based archival finding aids.

References

1. Cruikshank, K., Daniels, C., Meissner, D., Nelson, N.L., Shelstad, M.: How Do We Show You What We've Got? Access to Archival Collections in the Digital Age. Journal of the Association for History and Computing 8(2) (2005)
2. Doerr, M.: The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Magazine 24(3), 75–92 (2003)
3. Duff, W.M., Johnson, C.A.: A Virtual Expression of Need: An Analysis of E-mail Reference Questions. American Archivist 64(1), 43–60 (2001)
4. Duff, W.M., Johnson, C.A.: Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives. The Library Quarterly 72(4), 472–496 (2002)
5. Yakel, E.: Listening to Users. Archival Issues 26(2), 111–127 (2002)

Content-Based Image Retrieval in Digital Libraries of Art Images Utilizing Colour Semantics

Krassimira Ivanova

Institute of Mathematics and Informatics – BAS, Sofia, Bulgaria
kivanova@math.bas.bg

Abstract. The paper presents the architecture of experimental Content-Based Image Retrieval (CBIR) system APICAS ("Art Painting Image Colour Aesthetics and Semantics"). This system has been developed within a doctoral thesis which aims to provide a suite of specialized tools for CBIR within a digital library of art images. The high-level architecture suggested in this work takes OAIS as a basis and adds a designated layer to it allowing CBIR functions to be used both within ingest and access to the digital library.

Keywords: CBIR, OAIS, colour semantics, digital art.

1 Introduction

The development of specialized digital libraries (DL) for art images has to combine the traditional DL functionality with specialized image processing tools. Such tools can be used at ingest of digitized art objects as a means to enhance their metadata in automated way, or for access if the users would like to benefit from content-based image retrieval (CBIR) or other semantic-oriented tools. In this paper we are presenting architecture for a specialized art image DL which integrates general digital library functionality with designated CBIR tools. The suggested architecture had been implemented and the experience from this implementation informed this work.

2 Functional Requirements

The first step towards defining a suitable architecture for a CBIR system is to analyze the functional requirements it needs to meet. Our state-of the art review demonstrated that CBIR systems are developed most typically as specialized stand-alone applications or modules and are designed as such. This is a typical approach within an emerging domain but with the growing importance of image retrieval in the modern Web environment what becomes of special importance is how to develop modules for CBIR which could easily be integrated in digital repositories and web portals. This would require analyzing functional requirements for CBIR systems in the context of functional requirements within the current trends in digital archives. In order to address them, we will first present the high-level architecture of modern digital archives.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 515–518, 2011.

In 2002 the Consultative Committee for Space Data Systems prepared technical recommendations establishing a common framework of terms and concepts which comprise an Open Archival Information System (OAIS) [1] adopted later as the international standard ISO 14721:2003. This model can be successfully implemented as common framework with concretizations in application areas for so called GLAM (Galleries, Libraries, Archives, and Museums). The functional schema of OAIS contains six entities and related interfaces: **Ingest**, **Archival Storage**, **Data Management**, **Administration**, **Preservation Planning**, and **Access**. Within the context of such general digital archive architecture, CBIR-related implementations can be seen as a module which would best fit within the **Data Management** functional entity. However, it would also have influence on **Ingest** and more specifically on the structure of the submission information packages because the successful implementation of CBIR requires some specific data and metadata. CBIR also enriches the possibilities for delivery and will influence the **Access** functional entity which would accommodate more options for digital content discovery. This wider context is reflected in the architecture of a CBIR system called "Art Painting Image Colour Aesthetics and Semantics" (APICAS); this system is fine-tuned to the need of Information Retrieval (IR) in the area of digitized art collections. The specialized core part of the system accommodated the necessary specific instrumentarium in terms of algorithms and methods for IR; these are seen as specialized instances of **Data Management** tools. At the same time the special requirements for **Ingest** of specific data necessary for the IR components and the expanded **Access** possibilities are also highlighted.

3 APICAS Architecture

The software system APICAS was developed in order to supply appropriate environment for testing several kinds of visual and higher level features, connected with the colour presence and interaction between colours within art images [2][3][4][5]. In Figure 2 the architecture of proposed system is shown. The functional schema of APICAS follows OAIS excluding functional entities on administration and preservation planning. The Ingest functions in such experimental system are also very simplified, because the focus is on the extracting of visual metadata and analyzing received features. The main functions in APICAS are:

- data entry – establishing connections with image sources as well as supplying controlling textual metadata;
- feature extraction – producing automated metadata for image labelling;
- query interface – part of user-interface functions, connected with receiving of the tasks from the consumer. The image bank is used in order to select "an example" for searching images with greatest similarity. The metadata bank is used for constructing a "controlled vocabulary" for selecting desired feature(s);
- query processing – analysis of extracted metadata, their potential to meet user query for receiving images with specified colour harmonies or contrast or to be used for building artist practice profile or movement description;
- visualization – the other part of user-interface functions, connected with visualizing of received results. A variety of tools is used, such as image sets (whole images or patches), attribute data sets, distance files, graphics, knowledge analysis results, etc.

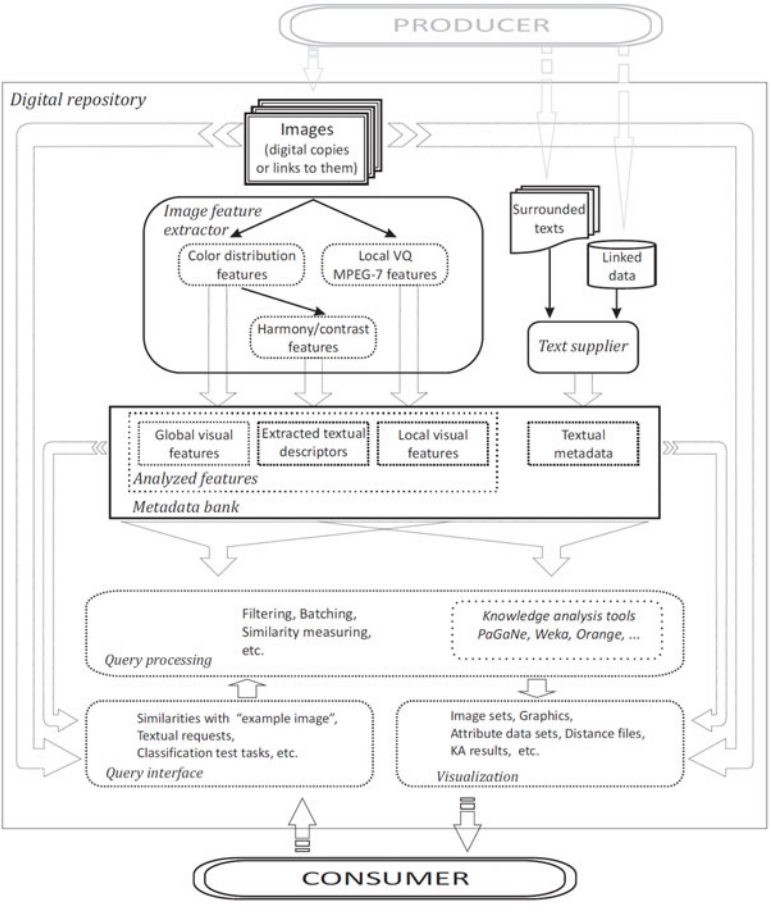


Fig. 1. APICAS architecture

The main goals of APICAS are in two-fold:

- to analyze the possibilities of defined harmonies and contrast features for narrowing the semantic gap;
- to investigate possibilities for finding regularities between these features that can be used as semantic profile of the art paintings.

The system is realized using CodeGear Delphi 2007 for Win32. As metadata storage space Arm 32, property of FOI Creative Ltd., is used. For obtaining the MPEG-7 descriptors APICAS refers to Multimedia Content Management System MILOS [6]. For obtaining the results of multidimensional scaling we used the open component-based data mining and machine learning software suite Orange [7]. As clustering algorithm "vcluster", a part of the CLUTO open source software package [8], is implemented in the system. As knowledge analysis environment we use the data mining environment PaGaNe [9], developed in the Institute of Mathematics and

Informatics, and especially Class-Association Rule classifier PGN, Association Rule Miner ArmSquare and implemented statistical analyzing tools for checking up our results and extracting regularities for artists' and movements' styles based on the extracted attributes. For comparing received results of PGN classifier we used Waikato Environment for Knowledge Analysis (Weka) [10].

4 Conclusion

We have proposed architecture of an experimental CBIR lab-system, aimed at analyzing different types of visual features, which strive to narrow the semantic and abstraction gap between low-level automatic visual extraction and high-level human expression. We have explained the structure and functionality of the software system "Art Painting Image Colour Aesthetics and Semantics" (APICAS). The vividness of proposed features will open the door for indexing and searching in paintings repositories, according to such characteristics of their content. The proposed features can be used as a step in the transition from Web 2.0 to Web 3.0.

Acknowledgements. This work was supported in part by Hasselt University under the Project R-1875 and by the Bulgarian National Science Fund under the Project D002-308.

References

1. OAIS: Reference Model for an Open Archival Information System (OAIS): Blue book. Consultative Committee for Space Data Systems, p. 148 (January 2002)
2. Ivanova, K., Stanchev, P., Dimitrov, B.: Analysis of the distributions of color characteristics in art painting images. Serdica Journal of Computing 2/2, 111–136 (2008)
3. Ivanova, K., Stanchev, P.: Color harmonies and contrasts search in art image collections. In: 1st Int. Conf. on Advances in Multimedia, Colmar, France, pp. 180–187 (2009)
4. Ivanova, K., Stanchev, P., Vanhoof, K.: Automatic tagging of art images with color harmonies and contrasts characteristics in art image collections. Int. Journal on Advances in Software 3(3&4), 474–484 (2010)
5. Ivanova, K., Stanchev, P., Velikova, E., Vanhoof, K., Depaire, B., Kannan, R., Mitov, I., Markov, K.: Features for art painting classification based on vector quantization of MPEG 7 descriptors. In: 2nd Int. Conf. ICDEM, India. LNCS (2010)
6. Amato, G., Gennaro, C., Rabitti, F., Savino, P.: Milos: A multimedia content management system for digital library applications. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 14–25. Springer, Heidelberg (2004)
7. Demsar, J., Zupan, B., Leban, G., Curk, T.: Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper, University of Ljubljana (2004)
8. Karypis, G.: CLUTO: A Clustering Toolkit Release 2.1.1. University of Minnesota, Department of Computer Science, Minneapolis, Technical Report: #02-017 (2003)
9. Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof, K., Stanchev, P.: PaGaNe – a classification machine learning system based on the multidimensional numbered information spaces. World Scientific Proceedings Series on Computer Engineering and Information Science (2), 279–286 (2009)
10. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

New Paradigm of Library Collaboration

Adam Sofronijevic

PhDc at University of Belgrade, Faculty of Philology
Studentski trg 3, Belgrade, Serbia
sofronijevic@unilib.bg.ac.rs

Abstract. The thesis entitled ''New Paradigm of Library Collaboration'' presents the case for the holistic approach to the issue of collaboration in a contemporary library. Patron needs and expectations in regards to collaboration, interactivity and ultimately participation are investigated in the specific area of changes in reading process. Collaboration between librarians and patrons and among librarians is discussed in regards to Library 2.0 and Enterprise 2.0 concepts. Based on the research results gathered in European libraries a new paradigm of library collaboration is presented as a must for an efficient library providing up-to-date services.

Keywords: Library collaboration, Web 2.0, Enterprise 2.0, Library 2.0, Digital libraries, European libraries, Library research, Enterprise 2.0 implementation in libraries, Reading 2.0, Solitary reader, Contemporary librarianship.

1 Introduction, Research Topic and Hypotheses

Based on the results of the original research and theoretical contemplation of subject meter the thesis presents the case for application of the holistic approach to issue of collaboration in libraries and a dire need for new paradigm of library collaboration.

The thesis investigates the fundamentals that allow for keeping operations of a library in line with its stakeholder needs and expectations. Researching the topic of changes in reading makes basis for understanding collaborative and participatory expectations and needs of patrons. Building on this Library 2.0 and Enterprise 2.0 related aspects of collaboration in a library are investigated. A holistic approach to these two sides of collaboration in libraries is presented and the need for new paradigm of library collaboration is explained.

The thesis encompasses three main subtopics. The first one concerns expectations of patrons that are used to a collaborative, interactive environment. The thesis deals with this subtopic by investigating changes in reading.

The second important subtopic of the thesis is collaboration between library employees and patrons described by the term Library 2.0. This subtopic is discussed on basis of the literary review of this area.

The third subtopic of the thesis is collaboration between library employees. Enterprise 2.0 describes the use of Web 2.0 tools and approaches by organizations in order to foster its internal functions. In libraries this concept describes the use of Web 2.0 tools by library employees for communication and collaboration with other employees.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 519–522, 2011.

The main research hypothesis is that for the understanding of processes that affect changes in libraries a new paradigm of library collaboration needs to be defined. The research hypothesis has twofold basis – patrons and library employees.

In the area of patron needs we predict that the rising importance of collaboration and interactivity in patron behavioural patterns form the basics for the ongoing changes. Libraries need to understand the nature of these continuing changes in order to implement the infrastructure that allows for continuing upgrade of its services. We predict that the expectations of patrons will remain within the boundaries of participatory institutions in whatever way the collaborative and interactive patterns of behaviour might evolve.

In the area of creation of basis for continuous efficient updating of library services we predict that implementation of possibilities for collaboration and interactivity among library staff is fundamental in allowing a library to cope with the fast changing, global environment. We predict that Enterprise 2.0 implementation in a library leads to better alignment of library services with patron needs and expectations.

2 Current Problems and Solutions

The literature review consists of three parts in order to give a sound starting point in discussing all three subtopic of the thesis. Literature review on reading encompasses considerable knowledge on this process gathered so far in regards to traditional reading process. Both the quantitative approach originating in psychology and qualitative approach originating in literature theories are depicted and main ideas presented so far important in regards to interactivity and collaboration elements are briefly described. Literature review on collaboration between users and library employees encompasses numerous research and theoretical papers on Library 2.0 that have appeared in the past. Literature review on collaboration between library employees consists mostly of relevant works regarding Enterprise 2.0 and its implementation originating in other industries. It also encompasses works that present solutions regarding employee collaboration in libraries.

Generally, the current situation in regards to collaboration in libraries is depicted having in mind three crucial components that build basis for collaboration in any organisation: underlying technology, social aspects and organisational aspects.

3 Approach and Contributions

The thesis aims at providing both theoretical and practical contribution to contemporary librarianship. The main theoretical contribution is in demonstrating the importance of holistic approach to collaboration in a contemporary library. Collaboration between patrons and library employees will be discussed in regards to the Library 2.0 concept on basis of previous works in this area. Collaboration between library employees will be discussed in regards to Enterprise 2.0 implementation on basis of original research results. Both aspects of collaboration will be then juxtaposed to each other in an attempt to create a holistic concept of collaboration in libraries.

An additional theoretical contribution will be given in the area of reading. The thesis present two new concepts: Reading 2.0 and Solitary reader. A new concept of

reading – Reading 2.0 is presented in order to describe a process that is not only passive and individualistic. Juxtaposed to Reading 2.0 the concept of the Solitary reader should provide us with a better understanding of reading process characteristics most affected by new possibilities in collaboration and interactivity.

The main practical contribution is in revealing the level of Enterprise 2.0 awareness and implementation in libraries in Europe. Some practical conclusions in regards to Enterprise 2.0 implementation processes will be given on basis of research results. The thesis advocates that working processes in a library might be improved in many ways thanks to the introduction of new technologies that forward collaboration and communication.

4 Preliminary Results and Further Research

The main focal point of the research is a library from Europe. Europe has been chosen because it encompasses libraries and patrons at various stages of advancement through the processes of technological and social innovation and because it was accessible for the researcher therefore allowing for gathering of representative set of data. Data from other regions are collected as well in order to provide points of reference.

Data on librarians themselves is collected directly via surveys and interviews. Data on patrons is collected indirectly from librarians.

The research is conducted in several phases and as of 2011 it is still ongoing. The first phase of the research was an online research conducted via Surveymonkey online survey tool. The questionnaire had two purposes. It was a data gathering tool and it was used to pinpoint libraries that are willing to share their experiences and knowledge during the second phase of the research. The three part online questionnaire was filled by representatives of 175 institutions.

The research so far has revealed that high levels of awareness of Web 2.0 applicability for business purposes in libraries exist. Implementation of Enterprise 2.0 is sketchy and ongoing. The second part of the questionnaire showed that infrastructure providing basis for fostering collaboration among patrons is not in place yet. E-book reader usage in libraries is registered but is not significant yet. The situation regarding collaboration in reading software is similar to lower general usage and awareness.

The second phase of the research is onsite research with data being gathered by means of structured interviews. So far 42 libraries in Europe have expressed the wish to participate in this phase of the research.

Third phase of research will comprise of virtual contacts with libraries that have already been visited in order to provide for fine tuning of research and gathering of additional data whose importance for the research topic might have been overlooked in previous phases.

5 Conclusions

The thesis is important not only for introducing into librarianship theory a novel problem of implementation of Enterprise 2.0 in a library, but also for providing some solutions that are discovered in the research process.

By defining the Reading 2.0 concept the thesis present an attempt in conceptualizing changes in reading process that follow various ongoing social and technological innovations. Abundant with ideas connecting existing important discourses in literary theory and elements of the Reading 2.0 concept the thesis present a rich source of ideas for theoretical consideration of issues posed by advances in technological and social innovations.

By describing phenomena that are touching various fields of expertise ranging from managerial and ICT issues to literary theory and psychology the thesis present case for a multidisciplinary approach to research and contemplation of digital libraries in the future.

References

1. Casey, M., Savastinuk, L.: Library 2.0: A Guide to Participatory Library Service. Information Today Inc., Medford (2007)
2. Crowder, R., Wagner, R.: The Psychology of Reading: An Introduction, 2nd edn. Oxford University Press, New York (1992)
3. Gadamer, H.G.: Truth and Method, 2nd rev. edn. Crossroad, New York (1989)
4. Gibson, E.J., Levin, H.: The Psychology of Reading. MIT Press, Cambridge (1975)
5. Granovetter, M.: The Strength of Weak Ties. American Journal of Sociology 78, 1360–1380 (1973)
6. Habib, M.: Toward Academic Library 2.0: Development and Application of a Library 2.0 Methodology. Master Thesis, University of North Carolina at Chapel Hill (2006), http://etd.ils.unc.edu/dspace/handle/1901/356
7. Iser, W.: The reading process: A phenomenological approach. New Literary History 3, 279–299 (1972)
8. Jauss, H.R.: The three horizons of Reading. In: Toward an Aesthetic of Reception. University of Minnesota Press, Minneapolis (1982)
9. Kendeou, P., van den Broek, P.: The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. Memory & Cognition 35, 1567–1577 (2007)
10. McAfee, A.: Enterprise 2.0: New Collaborative Tools for Your Organization's Toughest Challenges, Kindle edition. Harvard Business School Press, Boston (2009)

Visual Aesthetics of Websites: The Visceral Level of Perception and Its Influence on User Behaviour

Rita Strebe

Universität Regensburg,
Fakultät für Sprach-, Literatur- und Kulturwissenschaften,
Lehrstuhl für Informationswissenschaft,
Universitätsstr. 31,
93053 Regensburg, Germany
rita_strebe@web.de

Abstract. Website aesthetics has become an important research object in the domain of human-computer interaction during the last decade. Influences on acceptance and preference have been shown [1, 2]. The consideration of this quality aspect is also relevant for digital libraries as a possibility to appeal to the users on an emotional level. It is the aim of an empirical study to test the impact of the affectively effective aesthetics of websites on approach and avoidance behaviour. Thus the significance of this visceral level of perception is verified. In consequence of this fundamental research the applicability of affective reactions for the evaluation of website aesthetics could be further investigated.

Keywords: website aesthetics, user affect, affective priming, user behaviour, human computer interaction.

1 Introduction

The search interface of a digital library constitutes the first outward presentation of the library as perceived by the user. Being a mainly task oriented web application, the efficient support of the user in accomplishing her goals and thus a sound observation of the usability requirements is of prior importance. The influence of emotionally appealing quality aspects as joy of use and aesthetics on the overall experience of the user when interacting with the search interface should not be underestimated though. There is evidence that perceptions of the aesthetic quality and of the usability of an interface are interdependent [3]. Aesthetics of websites is the object of this investigation.

2 State of Research and Aims

A quality related approach within the research on user experience has the objective to identify quality aspects outside the ergonomic quality as assessed by usability evaluation instruments. The evaluation of the usability of an interactive product is oriented toward assessing its usefulness to support the user in accomplishing a certain

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 523–526, 2011.

task efficiently and effectively. Research on user experience brings into focus more task independent quality aspects as joy of use, originality and aesthetics, which appeal to the user on an emotional level of experience. In the past decade an increasing scientific interest in website aesthetics can be observed [1, 2, 3, 4, 5].

Norman established a differentiation of the aesthetic perception into a reflective and a visceral level. The reflective perception of aesthetics comprehends the cognitive and conscious interpretation and evaluation of an aesthetic object, while "the visceral level is pre-consciousness, pre-thought. It is appearance that matters here and first impressions are formed" [6]. Norman's thesis of such a pre-cognitive level of aesthetic perception is supported by several empirical findings. Evidence of an affective evaluation level which is effective independently of cognitive interpretation is mainly shown by experimental results on Mere Exposure Effect and Affective Priming Effect [7, 8]. Studies in the web context demonstrated, that the aesthetic impression is formed and consistent after a very short exposure time (50ms) and remains stable also after a longer exposure time, indicating that it is an affective evaluation level which influences the aesthetic perception of websites in a considerable way [4, 5].

It is the aim of this research project to test the significance of the visceral level of aesthetic perception of websites as a quality aspect. This is achieved by an investigation of approach and avoidance behaviour of test persons while interacting with aesthetically differing websites in a laboratory experiment. The results of this fundamental research may open up possibilities to apply affective reactions for the evaluation of website aesthetics.

3 Research Hypotheses

For the investigation of behaviour towards an object, the motivational direction and its valence are of importance. According to the valenced motivational direction, behaviour can be differentiated in approach and avoidance behaviour. Approach behaviour is a behavioural direction toward a positively valenced object. Avoidance behaviour is a behavioural direction away from a negatively valenced object [9].

Hypotheses for the planned investigation are accordingly:

H1: On a high level, the viscerally perceived visual aesthetics of websites effects approach behaviour.
H2: On a low level, the viscerally perceived visual aesthetics of websites effects avoidance behaviour.

4 Experimental Design

A one factorial design is chosen with aesthetics as independent variable. The viscerally perceived visual aesthetics of websites is set on two extreme levels and one intermediate level. Concerning the dependent variables, a multivariate design is established. The criterion for the selection of these variables is the applicability of their values for measuring approach and avoidance behaviour on different behavioural levels. The dependent variables are:

- dwell time on a website
- site penetration (number of visited pages)
- mean intensity of positive emotional deflections
- mean intensity of negative emotional deflections
- mean duration of eye fixations

To fulfil the ceteris paribus precondition, several more influencing factors on the behaviour towards the websites must be held constant as confounders:
- website usability
- topical content
- size of the interaction space

5 Selection of Stimulus Material

Mainly three requirements make an efficient selection of the stimulus websites necessary:
- It is necessary to select websites that differ significantly in their viscerally perceived aesthetic quality.
- The stimulus websites must be selected so as to control the confounders topical content, usability and size of the interaction space.
- The selection of websites as to their aesthetic quality is carried out on the basis of screenshots of the homepages while in the main investigation test subjects interact with the whole websites. To adapt the stimulus space, the stimulus websites must conform to further criteria as for example consistency of the layout.

The selection of stimulus websites is carried out in a five step selection process:
1. Determining a consistent topical content
2. Control of the size of the interaction space
3. Approximation of the stimulus space between preliminary and main investigation
4. Selection of websites with differing aesthetic quality through affective priming
   This experimental method allows for the assessment of the polarity of affective reactions to the visual quality of websites independent of a reflective evaluation.
5. Control of the website usability

6 Main Investigation

Each test subject interacts with four aesthetic, middle-rate aesthetic and not aesthetic websites respectively. For the interaction no tasks are preset. The course of the investigation is conceptualised in a way that allows the test subject to browse the different websites as realistically as possible so that approach and avoidance behaviour can express themselves undisturbed. The sequence of the websites is randomized for each session.

For data collection dwell time and site penetration will be collected through a screen recording of the sessions. Mean intensity of positive and negative emotional deflections are captured by measurement of the electromyographic activity of facial

muscles. Several investigations found out that a certain facial muscle over the brow (corrugator supercilii) shows increased activity with negative emotional polarity. Another facial muscle at the cheek (zygomaticus major) shows increased activity with positive emotional polarity [10, 11]. The mean duration of eye fixations is collected through eye tracking hardware and software.

For data analysis one-sided statistical tests are conducted on a significance level of 0,05 to verify the difference between the means of the dependent variables for aesthetic and middle-rate aesthetic websites. Thus the approach hypothesis is tested. The difference between the means of the dependent variables for not aesthetic and middle-rate aesthetic websites are verified to test the avoidance hypothesis.

References

1. Schenkman, B.N., Jönsson, F.U.: Aesthetics and Preferences of Web Pages. Behav. Inform. Technol. 19, 367–377 (2000)
2. Van der Heijden, H.: Factors Influencing the Usage of Websites: the Case of a Generic Portal in the Netherlands. In: 14th Bled Electronic Commerce Conference, pp. 174–185 (2001)
3. Tractinsky, N., Katz, A.S., Ikar, D.: What is beautiful is usable. Interact. Comput. 13, 127–145 (2000)
4. Lindgaard, G.: Attention Web Designer: You have 50 milliseconds to Make a Good First Impression. Behav. Inf. Technol. 25, 115–126 (2006)
5. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., Sharfi, T.: Evaluating the consistency of immediate aesthetic perceptions of web pages. Int. J. Hum.-Comput. St. 64, 1071–1083 (2006)
6. Norman, D.A.: Emotional Design: Why we Love (or Hate) Everyday Things. Basic Books, New York (2004)
7. Kunst-Wilson, W.R., Zajonc, R.B.: Affective Discrimination of Stimuli that cannot be Recognized. Science 107, 557–558 (1980)
8. Murphy, S.T., Zajonc, R.B.: Affect, cognition, and awareness: affective priming with optimal and suboptimal stimulus exposures. J. Pers. Soc. Psychol. 64, 723–739 (1993)
9. Elliot, A.G., Covington, M.V.: Approach and avoidance motivation. Educ. Psychol. Rev. 13, 73–92 (2001)
10. Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O.: Looking at pictures: affective, facial, visceral, and behavioral reactions. Psycholphysiology 30, 261–273 (1993)
11. Brown, S., Schwartz, G.E.: Relationships between facial electromyography and subjective experience during affective imagery. Biol. Psychol. 11, 49–62 (1980)

Revealing Digital Documents

Jakob Voß

Verbundzentrale des GBV (VZG), Göttingen, Germany
jakob.voss@gbv.de

Abstract. The research project aims at revealing common patterns that are used in data, independent from the particular technology in which the data is available. A better understanding of data patterns will not only help to better capture singular characteristics of data by metadata, but will also recover intended structures of digital objects.[1]

1 Introduction

Both documents and descriptions of documents increasingly exist as digital objects, basically as streams of bits, abstracted from any storage medium and location. While traditional concepts such as 'document', 'page', 'edition', and 'copy' blur, forms such as 'files', 'records', and 'objects', are rather different views on the same thing, than inherent properties of a digital document. These properties depend on context both for documents [3] and metadata [4], but you still need to actually look at data at some level of description, especially if context and formats an not fully known. A deeper look at data is required, to reveal how digital documents are actually structured and described. The question should not be answered by simply pointing to concrete technologies and formats, which are subject to rapid change and obsolescence, but at a more fundamental level. The main hypothesis is that all methods to structure and describe data share common patterns, independent from technology and level of description.

2 Background and Method

Three general concepts of data can be identified [2]: data as hard numbers, data as observations, and data as bits. Current "data science" mainly uses the first two connotations and applies statistical methods of data analysis to large data sets. This research commits to the third connotation, found in computer science and in library and information science. Both deal with information rather than data and mainly limit data research on issues of performance and preservation. The problem that I want to tackle more depends on the the inherent complexity of data, independent from its size and applications. Metadata research has shown [4, 6] that there is no silver bullet in data description but numerous ways to describe the same object by data, and the same data can describe different

[1] An extended version of this paper is available at http://arxiv.org/abs/1105.5832.

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 527–530, 2011.

things. Long-term preservation provides two strategies: either you emulate the environment of digital objects or you must regularly migrate them to new formats. Both strategies require good metadata, which themselves become subject of preservation, so documents may get buried in nested layers of metadata. In practice manual work is needed because context and function are not fully known or creators of data just do not comply to assumed standards.

A pattern describes "a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem" [1]. Such descriptions of good design practice have been adopted in software engineering [5] but less in the general design of data. Existing descriptions are limited to particular languages and formalisms. This hides general data patterns, independent from a particular encoding, and it conceals blind spots and weaknesses of a chosen formalism. For instance, nesting and order of elements in an XML document can be chosen with intent but also arbitrary just because documents must be ordered trees. To reveal patterns in data we cannot rely only on official descriptions and specifications. Existing approaches are either normative, or empirical but limited to one level of description which can then be analyzed with methods of data mining and machine learning. In contrast, I use a phenomenological research method that includes all aspects of data structuring and description: technical standards that specify data, software that shapes data, and how data is actually used by people. The phenomenological method views data as social artifacts, that cannot be described from an absolute, objective point of view. Instead data are studied as "'phenomena': appearances of things, or things as they appear in our experience" [7]. The main analysis consists of a review of diverse methods and systems for structuring and describing data, from simple character encodings to abstract data modeling languages.

3 Preliminary Results

An in-depth analysis of existing methods identified six non-exclusive groups: *character and number encodings* to express data, *identifiers* and *query languages* to identify data, *file systems* and *databases* to store data, *data structuring languages* and *markup languages* to structure data, *schema languages* to define and constrain data, and *conceptual modeling languages* to abstract and describe data. Existing methods are rarely discussed together as general structures with data as their common domain. Instead a strong focus on basic technologies like SQL, XML, and RDF is found. On a closer look, these technologies do not provide constant models but ocurr as slightly differing variants. A gap between research and practice is found, and confusion originates from differences between syntaxes, implementations, specifications, and models. It is shown how popular instances from the typology above each highlight a specific set of patterns and make other patterns less visible or more difficult to apply (see figure 1 for an example from a catalog of data patterns). Higher-level descriptions like schemas and modeling languages are no exception. Patterns and levels of abstraction often overlap and structuring methods can be used against their original purpose.

name
sequence pattern
idea
strictly order multiple objects, one after another
context
a collection of multiple objects
motivation
sequences are a natural method to model one-dimensional phenomena, for instance sequences of events in time. As digital storage is structured as sequence of bits, sequences seem to be the natural form of data and counterexamples, such as formal diagrams and visual programming languages, are often not considered as data.
implementations
• If objects have a known size, they can be directly concatenated. If objects have same size, this results in the array pattern.
• The separator pattern can be used to separate each object from its successor object. To distinguish objects and separators, this implies the forbidden objects pattern. If separators may occur directly after each other, this may also imply the empty object pattern.
• You can link one object to its successor with an identifier. To avoid link structures that result in other patterns (tree, graph, ...) additional constraints must apply.
• If objects have consecutive positions, a sequence is implied by their order.
examples
• string of ASCII characters (array)
• string of Unicode characters in UTF-8 (each character has known size)
• '`Kernighan and Ritchie`' (sequence with '` and `' as separator)
• $extract \rightarrow transform$, $transform \rightarrow load$, (sequence of linked steps)
counter examples
files in a file system, records in a database table, any unordered collection
problems
empty sequences and sequences of only one element are difficult to spot, like in other collection patterns.
similar patterns
without context, sequences are difficult to distinguish from other collection patterns. Many implementations of other patterns use sequences on a lower level.
implied patterns
position pattern
specialized patterns
array, ordered set, ring

Fig. 1. Example of a pattern description. Pattern names are underlined.

Typical examples include; the creation of dummy values for non-existing mandatory elements and the use of separators to add lists to non-repeatable fields. It appears that in practice it is not obvious which properties of data are intended and which arise as artifacts from the constraints of a given modeling language.

4 Evaluation and Application

The collection of patterns found during analysis can be evaluated by applying them to existing sets of data. Preliminary results confirm that common patterns (for instance identifiers, repeatability, grouping, sequences and ordering) are used on all levels of data description in different variants, implicitely and explicitly. Some patterns are already recognized, but it lacks a more systematic view, independent from the constraints of particular technologies. An examination of pattern that are actually applied to data shows that many description methods result in other structures than originally intended. It is unlikely that one single technology like XML or RDF will provide the final metadata tool. For this reason it is important to identify and apply patterns for both the creation of data and its consumption, especially in digital libraries. As the selection of patterns is a creative act of design, their recognition will to some degree free data designers from apologies and unquestioned habits that are justified as enforced by natural needs or technical requirements. Knowledge of general data patterns can help to reveal concealed structures in digital documents. Such retrospective analysis of incompletely defined or unknown data could be named "data archeology" and be located in the humanities, as it involves study of the cultural context of data creation and usage. Data patterns provide a contribution to intellectual data analysis. This analysis is needed to underpin and interpret algorithmic data analysis, which cannot reveal meaning of data as part of social practice. A general understanding of data and data patterns, as provided with this research, can therefore help libraries at least as much as the current understanding of physical publication types and materials.

References

1. Alexander, C., Ishikawa, S., Silverstein, M.: A Pattern Language: Towns, Buildings, Construction. Oxford University Press, Oxford (1977)
2. Ballsun-Stanton, B.: Asking about data: Experimental philosophy of information technology. In: Proc. of the 5th ICCIT, pp. 119–124 (2010)
3. Buckland, M.: What is a "digital document"? Doc. Num. 2(2), 221–230 (1998)
4. Coyle, K.: Understanding the semantic web: Bibliographic data and metadata. Library Technology Reports 46(1) (2010)
5. Gamma, E., Helm, R., Johnson, R., Vlissides, J.M.: Design Patterns: Elements of Reusable Object-Oriented Software, 1st edn. Addison-Wesley, Reading (1994)
6. Kent, W.: Data and Reality. Basic assumptions in data processing reconsidered. North-Holland, Amsterdam (1978)
7. Smith, D.W.: Phenomenology. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Summer 2009 edn. (June 2009)

Designing Highly Engaging eBook Experiences for Kids

Luca Colombo

Faculty of Informatics, University of Lugano (USI)
via Buffi 13, Ch-6904, Lugano, CH
luca.colombo@usi.ch

1 Introduction: Research Topic and Hypothesis

The HEBE (Highly Engaging eBook Experiences) project aims to explore how children can be involved into the design and evaluation of novel eBook interfaces in order to make the reading experience more engaging to younger audience.

The main aim of this research, in a nutshell, *will be to design a new concept of electronic book that enables the reader to "get lost" in the book, namely to enable a new, highly engaging, reading experience.*

The main hypothesis of the study is that *in order to make reading experience for children pleasurable and engaging, a new eBook model is needed, and children have to co-participate in its design.*

2 Research Context

The novelty that eBooks introduced for the first time ever is the possibility to engage the reader in an immersive reading of electronic texts. Immersive reading is a sort of deep reading that is commonly related to novels and that requires a significant investment of time and concentration.

But, children books are very different from those designed for adults. They are made to be more interactive and engaging. This way reading is just part of the whole *reading experience*. Those books offer a good example of creativity and variety, in terms of both content and presentation, which is hard to turn into electronic format.

3 Planned Approach

3.1 Engaging Reading Experience: A Theoretical Framework

In recent years HCI research is shifting its focus from the investigation of work-related issues (i.e. efficiency and efficacy of an application) to the study of user *experience*. ISO 9241-210 [15] defines user experience as "a person's perceptions and responses that result from the use or anticipated use of a product, system or service".

If we look at reading experience, we could affirm that it is engaging when we "get lost in a book", when we reach a state of deep concentration on the book we are reading, so much that we are not completely aware of what surrounds us and time flies without us noticing it. This happens more often when a reader is led to read by intrinsic motivations [1]. After looking at intrinsic motivation construct and some

S. Gradmann et al. (Eds.): TPDL 2011, LNCS 6966, pp. 531–534, 2011.

traits of being-lost-in-a-book experience, we noticed that these concepts fit together really well with *flow theory* [2].

Flow is defined as a mental state of deeply involvement / intense engagement in a certain activity, where most of a person's attention resources are devoted to accomplish that activity. Two key concepts of flow theory are: the balancing between the challenges and skills and the "autotelic experience", namely a self-contained activity that is done not with the expectation of some future benefit, but because the doing itself is the reward [2].

Researchers already investigated reading experience using flow model [3] or pointed out similarities between Csikszentmihalyi's *flow theory* and Nell's work on reading for pleasure [4, 5]. Moreover *flow theory* has been used also in some studies about human-computer interaction [6, 7] as a framework for modelling mental states like enjoyment, engagement, and pleasure.

3.2 Research Method and Evaluation Plan

The expected duration for the HEBE project is three years, each of which will roughly correspond to a different working phase.

Phase 1: literature review and "state of the art" analysis. Right now we are performing a critical review of existing eBook interfaces, reviewing scientific literature on topics related to Child Computer Interaction, building a bookshelf of children's novels, involving various stakeholders (parents, educators, librarians, etc.).

Our study starts from an extensive user study looking at how children, of different ages, interact with different types of book (traditional paper books and their electronic counterpart), in different environments (to include school, home and library), this will enable us to get a better understanding of users' needs.

Phase 2: formative evaluation and prototyping. In this phase children will be actively involved in the design of innovative eBook (intended as container) prototypes through *cooperative inquiry* approach [8]. Cooperative inquiry is a method to develop new technologies that includes three crucial aspects derived from user-centered approach: a multidisciplinary partnership with children; research on the field that allows a better understanding of context and activities; iterative prototyping (both low-tech and high-tech).

Phase 3: summative evaluation and methodological verification. The final stage of the research will consist in a summative evaluation of the prototype we will obtain from the previous stages. With this evaluation we will try to assess the overall reading experience and the level of engagement of the reader.

We believe Experience Sampling Method (ESM) to be the most promising method that could be adapted to gauge the engagement of the reader. ESM is gaining in popularity in the field of human computer interaction [9] and has been already used extensively in the research about flow [2] but also for the evaluation of UbiComp applications [10].

3.3 Main Challenges

The main challenge of this research is to build an effective evaluation technique to be used with children starting from the theoretical framework previously shown.

Major concerns when working with children are that they are more prone to bias answers when they respond to surveys [11]. Moreover, many ethical (and legal) constraints exist (i.e. privacy, consent, etc.) [11]. Finally, we will have to address the conceptual vagueness of flow, which, until now, has led to inconsistent operationalization of flow construct in the empirical work [12].

4 Early Outcomes

4.1 Building a (Good) Bookshelf

Ludic reading or reading for pleasure [5, 13] refers to the reading that we to do of our own free will, anticipating the satisfaction that we will get from the act of reading [1].

According to Nell's motivational flow chart of the antecedents and consequences of reading for pleasure [13], one of the three prerequisites needed in order to ensure that reading can be a pleasurable activity, is the correct book selection (the others two are readers' skills and readers' intrinsic motivations).

In order to have a good bookshelf to be used in the evaluation phases of the experiment, we have chosen books (or eBooks) most lent in libraries and best-sold in bookstores over the Italian territory during the last three years. We further refined our selection looking at readers' and stakeholders' (teachers, librarians, booksellers, parents, etc.) opinions.

4.2 "On the Field" Observation

From first observations on how children interact with books, we obtained some initial cues for our project. It seems that book illustrations, physical features (e.g. number of pages, font size, etc.), social component (e.g. book comparison, following the "trend", imitation, etc.) have a key role in children's reading experience. We believe that a good eBook prototype should support most of the above-mentioned aspects. We will investigate "how" in the future phases of this project.

5 Conclusions and Future Work

Following the literature review and "state of the art" analysis we performed, and the early outcomes we obtained, we believe *flow theory* to be a promising approach to analyse engagement while reading. That is why we will use it as guidance for this project and it is likely we will build our summative evaluation method upon it.

Our future work will aim not only to obtain information on how to develop more engaging eBooks for young (and old) readers, but also to create new research methods (or validate existing) to use in the HCI field since "[…] UX research calls for new methods and approaches for designing and evaluating experience" [14].

References

1. Clark, C., Rumbold, K.: Reading for Pleasure: A Research Overview (2006)
2. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper Perennial (2008)

3. Mcquillan, J., Conde, G.: The conditions of flow in reading: two studies of optimal experience. Reading Psychology 17, 109–135 (1996)
4. Worthy, J., Moorman, M., Turner, M.: What Johnny Likes to Read Is Hard to Find in School. Reading Research Quarterly 34, 12–27 (1999)
5. Hill, B.: The Magic of Reading (1999)
6. Hoffman, D.L., Novak, T.P.: Flow Online: Lessons Learned and Future Prospects. Journal of Interactive Marketing 23, 23–34 (2009)
7. Pilke, E.: Flow experiences in information technology use. International Journal of Human-Computer Studies 61, 347–357 (2004)
8. Druin, A.: Cooperative inquiry: developing new technologies for children with children. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems the CHI is the Limit - CHI 1999, pp. 592–599. ACM Press, New York (1999)
9. Khan, V.J., Markopoulos, P., Eggen, B.: Features for the future Experience Sampling Tool. In: Mobile Living Labs 2009: Methods and Tools for Evaluation in the Wild, pp. 31–34 (2009)
10. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate ubicomp applications. IEEE Pervasive Computing 2, 24–31 (2003)
11. Markopoulos, P., Read, J.C., MacFarlane, S., Hoysniemi, J.: Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers. Morgan Kaufmann Publishers Inc., San Francisco (2008)
12. Finneran, C.M., Zhang, P.: Flow in computer-mediated environments: promises and challenges. Communications of the Association for Information Systems 15, 82–101 (2005)
13. Nell, V.: Lost in a Book: The Psychology of Reading for Pleasure. Yale University Press, New Haven (1988)
14. Bargas-Avila, J.A., Hornbæk, K.: Old wine in new bottles or novel challenges: a critical analysis of empirical studies of user experience. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI 2011, pp. 2689–2698. ACM Press, New York (2011)
15. ISO technical committee 159 (Ergonomics of human-system interaction): ISO 9241 Part 210: Human-centred design for interactive systems. International Organization for Standardization (ISO), Geneva, Switzerland (2010)

Author Index