Yannis Ioannidis Paolo Manghi Pasquale Pagano
(Editors)

Proceedings of the Third Workshop on Very Large Digital Libraries VLDL2010

A Workshop in conjunction with the European Conference on Digital Libraries 2010

Organized by the Institute of Information Science and Technology of the National Research Council (ISTI-CNR) Pisa, Italy

Held in Glasgow, Scotland (UK),
10th of September 2010

Editors

Yannis Ioannidis
Department of Informatics,
National and Kapodistrian University of Athens,
Athens, Greece

Paolo Manghi
Istituto di Scienza e Tecnologie dell'Informazione (ISTI),
Consiglio Nazionale delle Ricerche,
Pisa, Italy

Pasquale Pagano
Istituto di Scienza e Tecnologie dell'Informazione (ISTI),
Consiglio Nazionale delle Ricerche, Pisa, Italy

Preface

The implementation of modern Digital Libraries is more demanding than in the past. Information consumers are facing with the need to access ever growing, heterogeneous, possibly federated Information Spaces while information providers are interested in satisfying such needs by sharing rich and organised views over their information deluge. Because of their fundamental role of information production and dissemination vehicle, Digital Libraries are also expected to provide information society with functionalities and services that must be available 24/7 and guarantee the expected quality of service. This scenario leads to the development of Very Large Digital Libraries, which are very large in terms of the number of information objects and collections to be made available, users to be served and potentially distributed functionality/content resources needed to construct them. Research on VLDLs opens up novel and actual scenarios, where researchers have to confront with new foundational and system design challenges in a context having scalability, interoperability and sustainability as focal points. Authors and participants of the past editions of Very Large Digital Library Workshop, respectively at ECDL 2008 and ECDL 2009, have confirmed the importance of the topic and eagerly started investigating the foundations of this new and hot research field.

The goal of the Third Very Large Digital Library workshop is to prosecute such fertile discussions, hence to continue on providing researchers, practitioners and application developers with a forum fostering a constructive exchange among all key actors in the field of Very Large Digital Libraries.

Our sincere gratitude goes to all the people who have directly or indirectly made this event possible. Among these our colleagues at ISTI-CNR, Donatella Castelli, Leonardo Candela, and Costantino Thanos for their research inspiration, the members of the program committee, who devoted part of their precious time to the success of this workshop, and of course to all authors, whose passion and ideas are the real fuel of VLDL 2010.

Yannis Ioannidis, Paolo Manghi and Pasquale Pagano
Organizers and Editors of the third VLDL workshop

Program Committee – Reviewers

Stefan Gradman
Institut fr Bibliotheks und Informationswissenschaft,
Humboldt-Universitt zu Berlin,
Germany

Kat Hagedorn
OAIster Project,
University of Michigan Digital Library Production Service,
USA

Dean B. Krafft
National Science Digital Library Project,
Cornell Information Science,
USA

Yosi Mass
IBM Research Division,
Haifa Research Laboratory,
University Campus, Haifa, Israel

Glen Newton
CISTI Research, Natural Research Council
Canada

Pasquale Savino
Technical University of Crete (MUSIC/TUC),
Greece

Supporting Institutions

VLDL 2010 benfited from the support of the following organizations:

- Institute of Information Science and Technology of the Italian National Research Council (ISTI-CNR), Pisa, Italy
- Organization of the European Conference on Digital Libraries 2010

Table of Content

Invited talk

Europeana Data Model: on Integrating Heterogeneous Digital Library Sources
Carlo Meghini

Workshop Contributions

Motivations for Crowdsourcing in Building an Evaluation Platform for Searching Collections of Digitized Books
Gabriella Kazai

The Papyrus News Ontology - A Semantic Web Approach to Large News Archives Metadata
Nadzeya Kiyavitskaya, Akrivi Katifori, Giulio Paci, Giorgio Pedrazzi and Roberta Turra

Rethinking Fingerprint Evidence Through Integration of Very Large Digital Libraries
Nadia kozievitch, Ricardo da Silva Torres Torres, Edward Fox, Sung Hee Park, Nathan Short, Lynn Abbott, Supratik Misra and Michael Hsiao

Capturing and Analyzing User Behavior in Large Digital Libraries
Giorgi Gvianishvili, Jean Yves Le Meur, Tibor Simko, Jerome Caffaro, Ludmila Marian, Samuele Kaplun and Martin Rajman

The Kramerius System - Open Source Solution for Digital Libraries
Tomas Foltyn

Error Tolerant Large Scale FRBRization
Andreas Juffinger and Elisabeth Lex

Motivations for Crowdsourcing in Building an Evaluation Platform for Searching Collections of Digitized Books

Gabriella Kazai

Microsoft Research, United Kingdom
v-gabkaz@microsoft.com

Abstract. In this paper we explore the use of Amazon's Mechanical Turk (MTurk) service to aid in the creation of a test collection for the evaluation of information retrieval systems on a large collection of digitized books. The context of our work is the INEX Book Track, which aims to evaluate approaches for supporting users in reading, searching, and navigating the full texts of digitized books. Our specific focus is the evaluation of book search systems based on the Cranfield paradigm, which requires the construction of a test collection, comprised of a set of digitized books, a set of user queries (or topics), and relevance assessments. We review the Book Track's efforts in the past three years to create such a collection with the help of its participants and explore a new approach employing crowdsourcing techniques, employing MTurk workers to create the topics of the test collection. Our results show crowdsourcing to be a viable option that can easily generate a high volume of test topics, but topic quality can vary greatly, leading to a rejection rate of 37%.

1 Introduction

The INitiative for the Evaluation of XML retrieval (INEX) aims to facilitate the evaluation of XML retrieval systems. In XML retrieval, also referred to as focused retrieval, systems aim to return to the user relevant document parts, instead of whole documents. Like TREC, the most well known evaluation campaign in IR, INEX follows the Cranfield paradigm of evaluation which is based on test collections constructed for the purpose and sets of measures to report and compare performance on the test collection. Reflecting the specialisation of INEX on evaluating XML retrieval, the test collection consists of XML documents, test topics that may have structural constraints and relevance judgements that reflect which parts of the documents, i.e., XML elements, are relevant to the topics.

Prompted by the availability of large collections of digitized books resulting from various mass-digitization projects [2], such as the Million Book project[1] and the Google Books Library project[2], the Book Track was launched in 2007 as part of INEX [3,4,5]. The overall aim of the track is to promote inter-disciplinary

[1] http://www.ulib.org/
[2] http://books.google.com/

research investigating techniques for supporting users in reading, searching, and navigating the full texts of digitized books, as well as to provide a forum for the exchange of research ideas and contributions. Toward this goal, the track provides opportunities for exploring research questions around three main topics: 1) Information retrieval (IR) techniques for searching collections of digitized books, 2) Mechanisms to increase accessibility to the contents of digitized books, and 3) Users' interactions with eBooks and collections of digitized books.

In this paper, we focus on the first of these areas, and in particular, our goal is to explore the use of Amazon's Mechanical Turk (MTurk) service to aid in the creation of a suitable test collection for the evaluation of book search systems on a large collection of digitized books. This is motivated by the need to scale up the Cranfield method for constructing test collections where the significant effort required to create test topics and to collect relevance judgements is otherwise inhibiting. For example, the estimated effort that would have been required of a single participant of the INEX 2008 Book Track to create and judge a single topic was to spend 95 minutes a day for 33.3 days [6]. By harnessing the collective work of the crowds, crowdsourcing offers an increasingly popular alternative for gathering large amounts of data feasibly, at a relatively low cost and in a relatively short time [1]. We are interested in using crowdsourcing to contribute to the building of a test collection for the Book Track, which has thus far struggled to meet this requirement by relying on its participants' efforts alone.

Before we detail our MTurk experiments, we first review the efforts of the Book Track in the past three years to create the test topics with the help of its participants.

2 INEX Book Track Evaluation Setup

At INEX, both the test topics and the relevance judgements are contributed by the participants, who in exchange for their efforts gain access to the completed test collection, a valuable resource enabling the evaluation and comparison of the performances of their search systems. In this section, we briefly describe the setup of the INEX Book Track and review the topic creation efforts of its participants between 2007 and 2009.

2.1 Participating Organisations

Since 2007, the Book Track has been attracting a growing number of participants, but the number of actively engaging groups remained at a more or less constant level. In 2009, a total of 84 organisations registered for the track, compared with 54 in 2008, and 27 in 2007. Of those registered, 16 took part actively in 2009, compared with 15 in 2008, and 9 in 2007. Active participants are those who contribute test topics, runs, or relevance judgements.

In 2009, 7 groups contributed 16 topics with 37 topic aspects (sub-topics), 4 groups submitted runs, and 7 groups contributed relevance judgements. In 2008,

Table 1. Active INEX Book Track participants between 2007-2009

ID	Institute	Created topics	Runs	Judged topics
2009				
6	University of Amsterdam	2	6	7
7	Oslo University College	2	20	2
12	University of Granada			2
14	Uni. of California, Berkeley		9	
29	Indian Statistical Institute			1
41	University of Caen	2		
52	Kyungpook National Uni.	2		
54	Microsoft Research Cambridge	2		6
78	University of Waterloo	2	4	2
86	University of Lugano	4		
	Unkown (crowdsourced)			2
2008				
6	University of Amsterdam	3	10	8
7	Oslo University College			1
14	University of California, Berkeley	2	3	
17	University of Strathclyde			3
30	CSIR, Wuhan University	4		
31	Faculties of Management and Information Technologies, Skopje	4		
41	University of Caen	2		3
52	Kyungpook National University	4		1
54	Microsoft Research Cambridge	8		17
56	JustSystems Corporation	3		2
62	RMIT University	4	10	13
78	University of Waterloo	4	8	4
86	University of Lugano	2		2
2007				
2	University of California, Berkeley		4	
22	Doshisha University	1		
23	Kyungpook National University	1		
26	Dalian University of Technology	5		
28	University of Helsinki	2		
36	University of Amsterdam	3		
54	Microsoft Research, Cambridge	13		
55	University of Tampere	5		
92	Cairo Microsoft Innovation Center		13	

a total of 11 groups created topics, 4 groups submitted runs, and 10 groups contributed to the relevance assessments. In 2007, 7 groups contributed topics, 2 groups submitted runs, while no judgements were collected. Table 1 provides a summary of the active participants in the last 3 years.

2.2 The Book Corpus

The corpus contains 50,239 out-of-copyright books, digitized by Microsoft, totalling 400GB and containing over 17 million page XML elements. The corpus is made up of books of different genre, including history books, biographies, religious texts and teachings, reference works, encyclopedias, essays, novels, and poetry. 50,099 of the books also come with an associated MAchine-Readable Cataloging (MARC) record, which contains publication (author, title, etc.) and classification information. The basic XML structure of a typical book is a sequence of pages containing nested structures of regions, sections, lines, and words.

2.3 The Search Tasks under Evaluation

Focusing on IR challenges, two search tasks are investigated: 1) The Book Retrieval (BR) task, framed within the user task of building a reading list for a given topic of interest, aims at comparing traditional document retrieval methods with domain-specific techniques that exploit book-specific features, e.g., back-of-book index, library catalogue information, etc. 2) The Focused Book Search (FBS) task aims to test the value of applying focused retrieval approaches to books, where users expect to be pointed directly to relevant book parts.

The evaluation of both these tasks requires test topics and relevance judgements collected at the page and book levels.

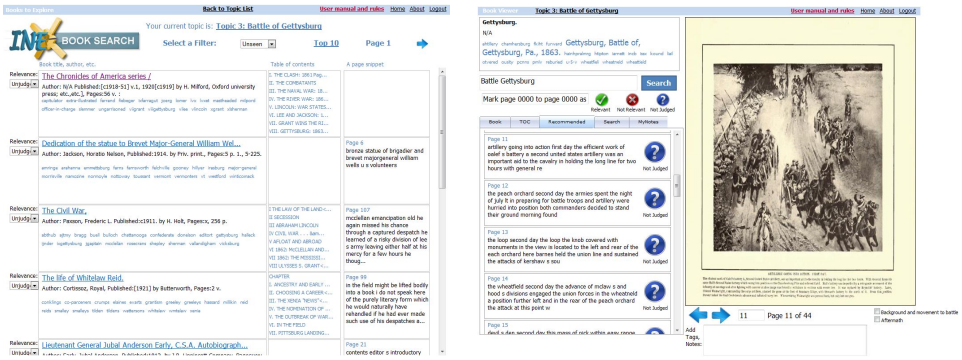


Fig. 1. Book Search System: showing a list of books (metadata, table of contents, and snippet from a recommended page) to be assessed for a given topic and the Book Viewer window with a list of pages to judge with respect to topic aspects.

2.4 Topics

A topic is a representation of a user's information need. In 2007 and 2008, topics had three parts:title, description and narrative, similarly to TREC topics. In

2009, topics could comprise of multiple aspects (sub-topics), where aspects are focused (narrow) needs with only a few expected relevant book pages.

Participants were asked to submit between 2-5 topics (varied from year to year), for which at least 2 but no more than 20 relevant books could be found in the corpus. The former condition aims to ensure that relevant content does exist for the topic in the corpus, while the latter condition is aims to ensure that the topic is not too easy. To aid participants in finding relevant books, an online Book Search System (http://www.booksearch.org.uk), developed at Microsoft Research Cambridge, was provided which allowed users to search, browse, read and annotate (mark pages relevant or add comments) the books in the test corpus. The system supports the creation of topics as well as the collection of relevance assessments [6], see Figure 1.

In 2007, 250 topics from the query log of Live Search Books were used in the BR task, and 30 topics (ID 1-30), created by participants, were used in the FBS task (then known as Page in Context task). In 2008, a total of 40 new topics (ID 31-70) were contributed by participants, which were used in both the BR and FBS tasks. In 2009, a total of 16 new topics (ID: 1-16), containing 37 aspects (median 2 per topic), were contributed by 7 groups (see Table 1). The 16 topics were used in the BR task, while the 37 topic aspects were used in the FBS task.

3 Quality of collected test topics

The creation of test topics, and in particular topics with structural constraints, by INEX participants has reportedly been suffering from artificial user needs due to the fact that participants are required to think up topics in order to contribute to the test collection [7]. While the book track only uses content-only topics (i.e., no structural conditions), badly formed or artificial topics may still get submitted. In our analysis, badly formed topics are those where some topic information is missing (e.g., no narrative). We consider topics artificial when the topic title contains a query from the Live Book Search query log, a sample of which was shared with participants. As it can be seen in Table 2, the quality of topics has been considerably improving over the past three years and it seems that participants are making a real effort to create a good quality test collection.

In addition to the above, one of the requirements of a good topic is that at least 2 but no more than 20 relevant books should exist in the top 100 search results obtained during topic creation (using the Book Search System). In Table 2, we report the number of topics with less than 2 or 10 relevant books in the full set of collected judgments. Topics with too many relevant books are those for which the ratio of relevant books and total judged books is over 60%. As it can be seen, having insufficient volume of relevance labels can render a large proportion of a test collection unusable for reliable evaluation: The total number of unusable topics is a direct result of topics with no judgements or topics with too few relevant results (less than 10 relevant books*). We note that there is high overlap between badly formed or artificial topics and topics that did not attract any relevance assessors and thus remained un-judged.

Table 2. Quality of topics created during INEX 2007-2009 by participants

Measure	2007	2008	2009
Total number of topics	30	40	16
Badly formed topics	6	0	0
Artificial information needs	13	3	1
Topics with no relevance labels	19	23	2
Too few relevant books (< 2)	5 of 11	3 of 17	0 of 14
Too few relevant books (< 10)*	8 of 11	10 of 17	3 of 14
Too many relevant books	4 of 11	6 of 17	3 of 14
Total unusable topics*	27	33	5

4 Crowdsourcing test topics

In preparation for the INEX 2010 campaign, we are experimenting with gathering topics both through Amazon's Mechanical Service and from the track participants. Our aim is to compare the quality of the collected topics and assess the feasibility of crowdsourcing topics (and relevance judgements later on).

To this end, we first redefined the search tasks, simplifying them in order to make topic creation for them suitable as a Human Intelligent Task (HIT) [1]. The two new INEX Book Track tasks are 'Prove It' and 'Best Books to Reference'. In the Prove It task systems need to find evidence in books that can be used to either confirm or refute a factual statement given as the topic. In the Best Books task systems need to return the most relevant books on the general subject area of the topic. To collect the test topics for the two tasks, we created the following two HITs:

- Facts in books HIT (Book HIT): "Your task is to find a general knowledge fact that you believe is true in a book available at http://booksearch.org.uk. Both the fact and the book must be in English. The fact should not be longer than a sentence. For example, the fact that 'The first Electric Railway in London was opened in 1890 and run between the stations: Bank and Stockwell' can be found on page 187 of the book titled 'West London' by George Bosworth". Workers were asked to record the factual statement they found, the URL of the book containing the fact, and the page number.
- Facts in books and Wikipedia HIT (Wiki HIT): "Your task is to find a general knowledge fact that appears BOTH in a Wikipedia article AND in a book available at http://booksearch.org.uk. You can start either by finding a fact on Wikipedia first, then locating the same fact in a book, or you can start by finding a fact in a book and then in Wikipedia. For example, the Wikipedia page on Beethoven's Symphony No. 3 claims that 'Beethoven dedicated the symphony to Napoleon, but when Napoleon proclaimed himself emperor, Beethoven tore up the title'. Page 144 of the book titled Beethoven by Romain Rolland describes this very fact". Workers needed to record the factual statement, the URL and page number of the book where the fact is found, as well as the Wikipedia article's URL.

We created 10 Wiki HITs, paying $0.25 per HIT, and issued two batches of Book HITs, with 50 HITs in each batch, paying $0.10 per HIT in the first batch and $0.20 in the second batch. All 10 Wiki HITs were completed within a day, while only 32 Fact HITs were completed in 11 days out of the first batch. The second batch of 50 Book HITs was completed fully in 14 days. The average time required per Book HIT was 8 minutes in the first batch and 7 minutes in the second batch (hourly rate of $0.73 and $1.63, respectively), while Wiki HITs took on average 11 minutes to complete (hourly rate of $1.31). These statistics suggest that workers found the Wikipedia task more interesting, despite it taking longer. However, as we show later, the attractiveness of a HIT does not guarantee good quality topics.

At the same time, INEX participants were asked to create 5 topics each, 2 of which had to contain factual statements that appears both in a book and in Wikipedia. A total of 25 topics were submitted by 5 groups. Of these, 16 facts appear both in books and in Wikipedia.

All collected topics were carefully reviewed and those judged suitable were selected into the set of test topics that is currently being used by the INEX Book Track. All topics contributed by INEX participants were selected, while filtering was necessary for topics created by MTurk workers. Out of the 10 Wiki HITs, only 4 topics were selected. Of the 32 Book HITs in the first batch, 18 were acceptable, while 36 were selected from the 50 Book HITs in the second batch. HITs were rejected for a number of reasons: the information given was simply an extract from a book, rather than a fact (20), the fact was too specialised (5), or nonsensical (5), the HIT had missing data (3), or the worker submitted the example given in the task description (1). Of the total 58 accepted HITs, 18 had to be modified, either to rephrase slightly or to correct a date or name, or to add additional information. The remaining 40 HITs were high quality and reflecting real interest or information need.

From the above, it seems clear that crowdsourcing provides a suitable way to scale up test collection construction: MTurk workers contributed 58 topics, while INEX participants created only 25 topics. However, the quality of crowdsourced topics varies greatly and thus requires extra effort to weed out unsuitable submissions. We note that selecting workers based on their approval rate had a positive effect on quality: batch 2 of the Book HITs required workers to have a HIT approval rate of 95%. In addition, paying workers more also shows correlation with the resulting quality.

5 Conclusions

The INEX Book Track, currently in its fourth year, has been attracting considerable interest but suffers from low active participation due to high costs in terms of required effort, including the need to provide test topics and relevance judgements. This year, we are experimenting with using MTurk to gather both topics and relevance judgements in addition to collecting these from the track participants. Our aim is to test the reliability of the crowdsourcing approach

so that in future years we can move the test collection creation completely to crowdsourcing.

In this paper, we summarised the last three years of the track and the challenges experienced in collecting test topics from participants, resulting in suboptimal quality topics that fail to attract relevance assessors, rendering the topics unusable in the evaluation. We described our approach for crowdsourcing topics on MTurk, which promises to be a more effective approach, leading to more realistic topics. However, the quality of crowdsourced topics varies greatly, requiring manual pruning: Out of the total 92 topics, 34 had to be rejected (37%). While higher pay and worker selection improve quality, the incentives in the crowdsourcing setup are no match to those at INEX, where participants directly benefit from the quality of their work in the created test collection. At the same time, the average time that workers invested (which was much greater than predicted) suggests that they generally had good intentions to do well in the task – behaviour also observed in [8]. 26 workers even provided extra comments.

The real test of the quality of the collected topics will, however, be decided later on once the relevance labels have been collected. We can then report on measures such as those shown in Table 2 and evaluate the feasibility of the crowdsourcing method for constructing the Book Track test collection. We are currently researching ways to ensure the quality of crowdsourced relevance labels, where manual inspection of the data is not viable.

References

1. Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.
2. Karen Coyle. Mass digitization of books. *Journal of Academic Librarianship*, 32(6):641–645, 2006.
3. Gabriella Kazai, Antoine Doucet, and Monica Landoni. Overview of the inex 2007 book track. In *Advances in Focused Retrieval, 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, 2007.
4. Gabriella Kazai, Antoine Doucet, and Monica Landoni. Overview of the inex 2008 book track. In Shlomo Geva, Jaap Kamps, and Andrew Trotman, editors, *INEX*, volume 5631 of *Lecture Notes in Computer Science*, pages 106–123. Springer, 2008.
5. Gabriella Kazai, Marijn Koolen Antoine Doucet, and Monica Landoni. Overview of the inex 2009 book track. In *Advances in Focused Retrieval, 8th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2009*, 2009.
6. Gabriella Kazai, Natasa Milic-Frayling, and Jamie Costello. Towards methods for the collective gathering and quality control of relevance assessments. In *SIGIR '09: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 2009.
7. A. Trotman. Wanted: Element retrieval users. In *Proceedings of the INEX 2005 Workshop on Element Retrieval Methodology, Second Edition*, pages 63–69, 2005.
8. Dongqing Zhu and Ben Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, 2010.

The Papyrus News Ontology – A Semantic Web Approach to Large News Archives Metadata

Nadzeya Kiyavitskaya[1], Akrivi Katifori[2], Giulio Paci[3], Giorgio Pedrazzi[3], Roberta Turra[3]

[1]Department of Information Engineering and Computer Science, University of Trento, via Sommarive 14, 38100 Trento, Italy
nadzeya@disi.unitn.eu
[2]Department of Informatics and Telecommunications, National and Kapodistrian Univeristy of Athens, Panepistimioupolis, Ilissia, 157 84, Athens, Greece
vivi@di.uoa.gr
[3]CINECA - Consorzio Interuniversitario, via Magnanelli 6/3, 40033, Casalecchio di Reno, Italy
{g.paci, g.pedrazzi, r.turra}@cineca.it

Abstract. Recent progress in digital library applications has created new possibilities for available archival content in the Web. Still, a large effort needs to be realized in order to render this content truly accessible and usable. To address this urgent need, we have been working towards developing a platform for storage, processing and semantic information retrieval in multimedia digital archives. This paper discusses a number of challenges encountered in designing the ontology which provides the archival content metadata of such a platform and the solutions proposed with a focus on scalability issues.

Keywords: digital libraries, ontologies, multimedia archives, scalability.

1 Introduction

Recent progress in the area of digital libraries (DL) has offered new ways of digitizing, organizing and presenting library material. Libraries and organization archives have started to digitize their material, either for internal use or for publishing it through the Web. The great variety of digitized content has resulted in new user needs and research targeting the development of new methodologies and tools.

The Papyrus project, started in 2008, attempts to address issues of information retrieval within this diverse and large DL content by providing a set of semantic web tools for content annotation and access. It intends to provide a dynamic DL 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 achieve this, the source content has to be 'understood', i.e., analysed and modelled according to a domain ontology. The user query also has to be 'understood' and analysed following a model of this different discipline. Correspondences must be then found between the model of the source content and the realm of the user knowledge. Finally, the results must be presented to the users in a useful and comprehensive manner according to their own 'model of understanding'.

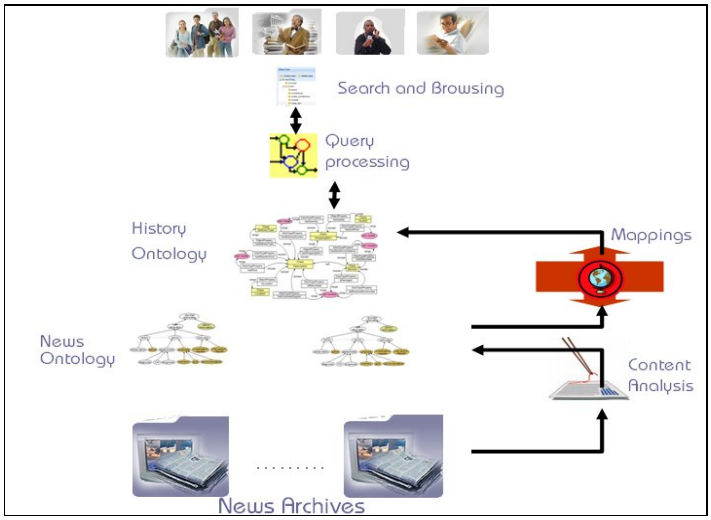


Figure 1. The Papyrus Platform

To realize this vision, Papyrus has applied and extended existing Semantic Web technologies. The Papyrus platform (Figure 1) is designed with two ontologies at its core [4], the History and News one, which model the History and News domains respectively. The two ontologies were created as extensions of existing standards with the cooperation of the corresponding domain experts, journalists and historians. In Papyrus, the news archives of Agence France Press[1] (AFP) and Deutsche Welle[2] are annotated in XML and stored in a relational database. The current working prototype is already available[3] and its preliminary evaluation is concluded.

The History ontology formalizes the domain of History of Science and Technology focusing on such phenomena as discipline formation, evolution, social and ethical aspects, and other so-called historiographical issues. This ontology was built on top of an ISO 21127:2006 standard CIDOC CRM for describing concepts and relationships used in cultural heritage domain. In contrast, the News ontology contains the constructs describing the details of a news item's provenience and storage, and basic semantic concepts. The News ontology is mapped to the History ontology by taking advantage of the historical research method to retrieve information on specific historical topics. The platform also offers a specialized web-base ontology browser [8] which, together with the keyword search and the mapping mechanisms of the platform, enables navigation from History ontology entities to News ontology entities and effective access to the archival material. Several Web tools were also developed for distributed multi-user ontology editing, creation of mappings between the two domains, and management of news content and analysis results [1].

One of the most vital issues that Papyrus had to address is scalability. All DL news items are automatically annotated with concepts in the News ontology using a method

[1] AFP website: www.afp.com
[2] Deutsche Welle website: www.dw-world.de
[3] Papyrus platform prototype: http://iris.atc.gr/CMS_Papyrus_1_1/

developed in the context of the Papyrus project [7]. The most relevant keywords in a news item are detected and connected to the most appropriate ontology classes based on a relatedness measure relying on Wikipedia knowledge. The automatic annotation allows managing a large number of news items. AFP news production, for example, is roughly 5000 new dispatches a day in six main languages (French, English, Spanish, German, Portuguese and Arabic). French and English are the main production with around 800-1000 dispatches a day. Those wires are about 320-350.000 news per year. Hence, the News ontology is expected to continuously grow with the terminology for accommodating the annotation needs of the news items. We elaborate on these challenges in constructing large-scale DLs and propose a number of solutions.

The rest of the paper presents the related work (Section 2), the News ontology (Section 3) and its semi-automatic creation method (Section 4). Section 5 discusses scalability issues and Section 6 concludes the paper.

2 Related Work

The Papyrus News ontology goes one step beyond other similar initiatives to create annotation frameworks for News content.

The **NEWS Ontology** [2] was developed in the context of the NEWS[4] project. It covers the main concepts required in the news domain. It is a lightweight RDFS ontology and provides the basic constructs for news item categorization and content annotation. Another similar approach is the New York Times Linked Open Data[5]. Our work complies with the Linked Data[6] vision for annotating and sharing data on the Web with the limitation that only a part of the news content we have been using will be available for public use due to the copyright protection.

NewsML[7] was designed by the IPTC[8] (International Press Telecommunications Council) to provide a media-independent, structural framework for multi-media news. A new major version of this standard, named **NewsML-G2**[9], was released in 2008. The **IPTC Subject News Codes**[10] are sets of topics to be assigned as metadata values to news objects like text, photographs, graphics, audio- and video assets, thus enabling a consistent coding of news metadata over the course of time. Using these codes is recommended by the IPTC for the classification of NewsML documents. This standard is used by major news providers like AFP, EBU[11] and Reuters Media[12].

The Papyrus News ontology extends the NewsML-G2 standard by organizing the standard and the IPTC Subject News Codes into an OWL ontology with a richer named entities and concepts structure.

---

[4] NEWS (News Engine Web Services) Home: http://www.news-project.com
[5] Linked Open Data of the New York Times: http://data.nytimes.com/
[6] Linked Data: http://linkeddata.org
[7] NewsML standard: http://www.newsml.org
[8] IPTC website: http://www.iptc.org
[9] http://www.iptc.org/cms/site/index.html?channel=CH0111
[10] http://www.iptc.org/NewsCodes/index.php
[11] European Broadcasting Union: www.ebu.ch
[12] Reuters website: www.reuters.com

3 The News Ontology

The News ontology [4] was developed within Papyrus in close cooperation with news professionals working in AFP and is intended to describe the structure and the semantics of the news content. The ontology was constructed based on the NewsML-G2 XML standard. For the needs of the Papyrus project, we integrated two different parts in the ontology: (a) the modeling of the format in which news items are produced by the main news agencies, i.e., the constructs adopted from NewsML-G2 (the presentation of this part of the ontology is omitted in the present paper; more information can be found in [1, 4]), and (b) the modeling of concepts present in the news items and relevant to the application domains, i.e., Biotechnology and Renewable Energy. These include named entities, concepts to accommodate domain-specific concepts, and instances. We further discuss the basic structure of the Papyrus extension of the ontology.

In the extended model (Figure 2), each news item is identified by its URI and can have a list of related topics that may contain: *themes* – IPTC categories to be respected by the news agencies when annotating their news content, as well as domain-dependent – and *terms*, such as *named entities* (like Person, Organization, Location), *concepts* (other entity types), or *slugs*, i.e., terms defined as relevant to the IPTC subjects. In turn, each term can be defined by a set of keywords. Thus, a news item has a rich set of metadata, for instance a theme "Cloning", a location "Seoul", an event "press-conference", a person "Hwang Woo-suk", and similar.

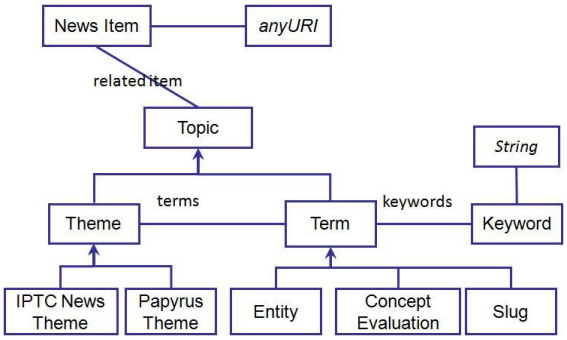


Figure 2. News ontology model for annotating news items. Arrows represent *is-a* relations and named arrows role ones

In more detail, *Topic*, *Theme* and *Term* are abstract concepts in this model and their underlying concepts are:

- *IPTCNewsTheme*. In the News ontology we adopted those IPTC categories that can be important for two application domains of Papyrus: biotechnology and renewable energy. To do so, the AFP experts manually selected a subset of IPTC topics that may contain information pertaining to either Renewable Energy or Biotechnology areas. As a result, 280 instances of this class have been included in the News ontology.
- *PapyrusTheme*. In order to represent more specific domain knowledge that is not represented by the IPTC categories, we created a new class,

PapyrusTheme. All domain-related topics have been represented as instances of this class, starting from the two main topics of interest: Renewable energy and Biotechnology, and then their subtopics, such as Cloning, Stem cells, Hydrogen energy, Environmental protection, and others. In order to support is-a relations between the instances, we exploited *skos:broader* and *skos:narrower* properties. Where relevant, we also linked IPTC news topics to one of Papyrus domains using *skos:sameas* property. So far the News ontology contains 32 Papyrus Themes.

- *Entity*. The News ontology was largely populated with varied types of named entities. The taxonomy of named entities extends the usual three classes – Organization, Location, Person (2,820 instances). The Location class of entities is represented by "GeoArea" concept, while Person and Organization are grouped under a more general concept called "Party". Apart from these common types of entities, we added the concepts of "Event", "Landscape", and "POI" (Point of interest) that includes, for instance, monuments.
- *ConceptEvaluation*. Instances of this concept are used to group several single keywords under one entity (e.g., "rotor blades", "rotor blade", "blades"). At the moment the ontology contains 6,930 ConceptEvaluation instances.
- *Slug*. This construct is inherited from the IPTC categorization, where each IPTCNewsTheme can be assigned one or more slugs, i.e., relevant terms. In total, 205 slug instances were selected given the two Papyrus domains.
- *Keyword*. Finally, the Keyword concept stores natural language expressions related to varied Term types. The total number of instances is around 30,000.

Thus, a (Papyrus or IPTC) theme instance can be related to a set of Entities, ConceptEvaluations or Slugs by means of "terms" relationship, where these are defined by sets of Keywords.

It is important to emphasize the multilingual nature of the Papyrus ontology, where most of the ontology instances have been assigned corresponding translations in three languages (English, French and German). The language is specified by the "xml:lang" attribute of a keyword's value, like for instance in this Location entity:

```
<Country rdf:ID="Country_00065">
  <keywords>
    <Keyword rdf:ID="Name_fr_00612">
      <value xml:lang="fr">Republique tcheque</value>
    </Keyword>
  </keywords>
  <keywords>
    <Keyword rdf:ID="Name_Country_00128">
      <value xml:lang="en">czech republic</value>
    </Keyword>
  </keywords>
</Country>
```

4 News ontology population method

The population effort has been undertaken for (a) the two domains that Papyrus focuses on, i.e., Biotechnology and biomedical technology and Renewable energy

with focus on wind power, and (b) three languages, i.e., English, French and German. Our ontology population method combines several different techniques.

The tool used for population of named entities is Stanford Named Entity Recognizer[13] [3]. An English model, trained on the CoNLL 2003 English data, was used to recognize 3 classes (Location, Organization, Person). This tool provides a general (arbitrary order) implementation of linear chain Conditional Random Field sequence models coupled with feature extractors for Named Entity Recognition. A named entity in a corpus can fall into more than one class depending on the context.

For the keyword extraction task, a two-step approach has been adopted: (1) terminological candidates are extracted by linguistic processors (part of speech tagging, phrase chunking), (2) and then terminological entries are filtered from the candidate list using statistical methods. Firstly, terminological candidates (aka multiword terms) were extracted from plain text using TreeTagger[14] [10] as NLP component. The final extraction of multiword terms was obtained using the shallow parsing procedure (phrase chunking) of TreeTagger and excluding the Named Entities already found. The process of keywords extraction has been repeated separately for the two main domains and for each subcategory. Secondly, the candidates list has been filtered using a measure (TF-IDF) to assess the relevance of a certain multiword term with respect to the whole corpus. Another measure RFR (Relative Frequency Ratio) has been added to assess how a multiword term is specific to a subcategory in respect of the use in its main domain. The idea is that multiword terms that occur relatively frequently in a subcategory compared to how frequently they occur in the general corpus are more likely to be good keywords. The relative frequency of a multiword term can be used as an initial filter.

In order to identify instances of ConceptEvaluation, which are basically represented by groups of similar keywords, we applied a technique based on the knowledge base of Wikipedia [7]. For each keyword the most appropriate Wikipedia page has been detected using a disambiguation process [6] that takes into account the subcategory context. If different keywords are linked to the same page, with a high disambiguation probability, they are grouped together as a unique concept.

5 Scalability Issues

To make the News ontology a useful tool for realistic digital archives applications that contain thousands of content items, we had to address several issues related to the scalability of our approach. When automatic tools are concerned, scalability has not proven to be an important issue, however we had to improve and extend existing techniques given the large size of our ontology. On the other hand, the points that required human intervention and work are the most problematic.

Firstly, the News ontology construction for digital archives of news items entails a number of scalability issues at several levels:

Domain change. Adding a new domain of archival content requires populating domain-dependent keywords and concepts. In this case, our automated method for

[13] Stanford NLP Group: http://nlp.stanford.edu/software/CRF-NER.shtml
[14] TreeTagger website: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

extraction of multiword terms based on TreeTagger and their grouping by means of Wikipedia can be reused.

Multilingualism. Instances population for new languages is not an easy task, as both domain-dependent keywords and named instances must be revised. For populating the ontology with German and French instances, the English keywords have been translated using the Google translator service. The automatic translation has been manually revised. The tool used for Named Entities is adaptable to other languages and to other classes if a manually tagged training corpus is available.

Ontology schema revision. When new requirements emerge, new concepts are often introduced in the ontology. Their population requires development of additional (semi-)automated tools. The solution depends on the nature of the proposed concept. For instance, in the course of Papyrus we were asked to add a concept of Event in the ontology, embracing such entities as conferences, meetings, social revolutions, scientific breakthroughs and others. In this case, we manually identified a small set of relevant instances straightforwardly from the user requirements. We also consider extending this set by using the recent ontology of events [5].

Another important issue related to the ontology size was the requirement to be able to view and use in the mapping creation process, not only the Papyrus Themes, also the ConceptEvaluation instances used by the content analysis for classification. These instances record a wide range of domain entities at a level of detail which is interesting to the digital archive user. However, only for the two Papyrus domains, this class contains 7000 instances. This number proved to be prohibitive for existing web-based ontology editors like Protégé. To address this issue in our own web ontology we applied paging techniques. As shown in Figure 3, the ConceptEvaluation instances are presented in 579 pages, which the user may browse without any delay.

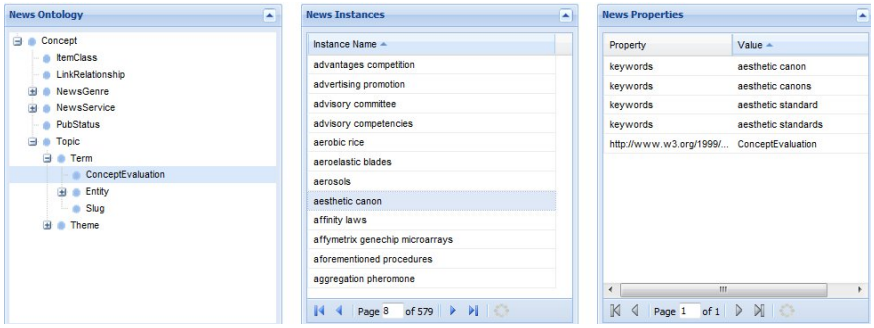


Figure 3. Part of the News ontology as it is presented in the Papyrus ontology browser. The ConceptEvaluation class is selected.

Mapping the News ontology entities to History ontology task is also a scalability related issue. These mappings are so far performed mostly manually, which is a very time-consuming task. For one of the News ontology classes only, Concept Evaluation, the user has to go through almost 7000 instances and try to create mappings to similar History ontology classes or instances. Although the preliminary evaluation results showed that the users could use the mapping tool efficiently, the number of instances involved still makes the task very time consuming.

We have been working towards developing an automated tool that may propose simple mapping candidates to the user, to be quickly revised. Advanced, intelligent mapping tools are needed to achieve greater automation of this process.

In general, the solutions adopted for addressing the scalability issues encountered within Papyrus can be reused in ontology population tasks of other applications.

6 Conclusions

This work presents the Papyrus News ontology for multimedia digital archives. We discussed the challenges of building and populating such an ontology and described the approaches used to address these challenges in the framework of the project. The proposed ontology represents the collaborated effort of experts from different areas, news professionals and computer scientists.

Our main contributions include: (a) an ontology schema which combines standard IPTC constructs for news content exchange and more 'semantic' constructs that allows for semantic information search, (b) a reusable method along with a toolset for ontology population that addresses scalability concerns.

Acknowledgements: This work was partially funded by the EU FP7 ICT framework. We gratefully acknowledge our partners from AFP and Deutsche Welle for their contribution in continuous revisions of the ontology and helpful discussions.

References

1. Bykau, S., Kiyavitskaya, N., Tsinaraki, C., Velegrakis, Y.: Bridging the Gap between Heterogeneous and Semantically Diverse Content of Different Disciplines. To be published in Proceedings of FlexDBIST-2010, Bilbao, Spain, September 2, 2010.
2. Fernandez-Garcia, N., Sanchez-Fernandez, L.: Building an Ontology for NEWS Applications. In Poster Session of the 3rd Int. Semantic Web Conf., ISWC, 2004.
3. Finkel J. R., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In Proc. of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), 2005.
4. Kiyavitskaya, N.: Documentation on Papyrus ontologies, Technical report available at http://www.ict-papyrus.eu/files/Documentation%20on%20Papyrus%20Ontologies.pdf
5. Lösch, U., Nikitina, N.: The newsEvents Ontology - An Ontology for Describing Business Events. In: Workshop on Ontology Design Patterns, ISWC, 2009.
6. Milne, D., Witten, I. H.: Learning to link with Wikipedia. In Proc. of CIKM'08, pp. 509–518, New York, NY, USA. ACM, 2008.
7. Paci, G., Pedrazzi, G., Turra, R.: Wikipedia-based approach for linking ontology concepts to their realisations in text", In Proc. of LREC'2010, Malta, May 17-23, 2010.
8. 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 Proc. of WIAMIS 2010, Desenzano del Garda, Italy.
9. Rizzolo, F., Velegrakis, Y., Mylopoulos, J., Bykau, S.: Modeling Concept Evolution: A Historical Perspective. In Proc. of ER 2009, pp. 331-345.
10. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In Proc. of the Int. Conf. on New Methods in Language Processing, 1994, pp. 44–49.

Rethinking Fingerprint Evidence Through Integration of Very Large Digital Libraries

Nádia P. Kozievitch[1], Ricardo da S. Torres[1], Sung Hee Park[2], Edward A. Fox[2], Nathan Short[3], A. Lynn Abbott[3], Supratik Misra[3], Michael Hsiao[3]

[1] Institute of Computing, University of Campinas, Campinas, SP, Brazil
{nadiapk,rtorres}@ic.unicamp.br
[2] Department of Computer Science
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
{shpark,fox}@vt.edu
[3] Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
{nshort2,abbott,supratik,hsiao}@vt.edu

Abstract. Fingerprints play a key role in biometrics and forensic science because of their uniqueness. Essential is contextual integration of fingerprint evidence from different sources, which involves composing, reusing, and aggregating a large amount of information. Thus, this paper (1) describes different types of fingerprint information from a digital library perspective; (2) investigates compound object concepts as used in connection with fingerprints; and (3) presents a preliminary integration of very large fingerprint digital libraries.

1 Introduction

Fingerprints have been used for identification from the early 1900s. The patterns formed by the ridges are important since they already are formed in the fetus by the fourth month of pregnancy and do not change until death. These patterns cannot be altered, except by accident, mutilation, or very serious skin disease, as they are formed in deep layers of the dermis. The skin consists of two main layers: the outer skin or epidermis, and the inner or true skin, known as the dermis.

The common friction ridge patterns – loops, whorls, and arches – impart class characteristics to a fingerprint [1], pre-aligning algorithms according to these singularities. This is similar to large image retrieval [2] systems, where there is a pre-analysis of quality, direction, ridge flow, angles, etc.

Our contribution is the analysis of fingerprint related activities, unifying different domains, using a digital library (actually, 4 DLs) and compound object (CO) perspective. Those aware of law enforcement activities will know of the first type of DL (DL1), associated with databases of stored fingerprints. Another consideration relates to our BAE Systems funded project to create training materials for fingerprint examiners, which leads to a second type of DL (DL2). A third type of DL relates to the evidence and data describing a crime scene (DL3).

A fourth type of DL relates to our NIJ funded research studies supporting experimentation with fingerprint image analysis techniques, quality measures, and matching methods (DL4). Combining these four into an integrated DL, where compound objects allow us to work across these DLs (see Fig. 1), yields a very interesting and very large DL.

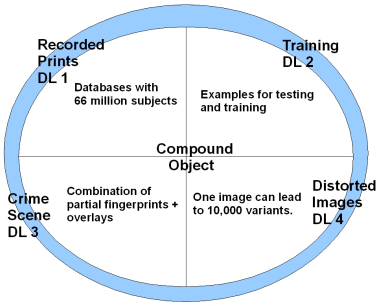


Fig. 1. The integration of fingerprint digital libraries.

In DL1, information is used to identify a person. DL2 has a different purpose: to educate and train users. In DL3, images are used for matching or excluding individuals. In DL4, the focus is on algorithms, varying parameters. Through our integration, digital libraries unify four different communities, allowing each one to see different perspectives, and explore the system as a whole, or focus in a determined area. In addition, we can take advantage of digital library services (e.g., browsing and searching), formalisms, and preservation solutions.

We plan to use compound objects (COs) [3] to facilitate aggregation abstraction, embracing components from different domains, and unifying them with a single concept. COs also can help us achieve benefits arising from the script concept of Schank [4], where components in a CO have the same behavior, or respect the same rules. Finally, solutions to some of the "very large" issues in digital libraries result from using COs, e.g., when specific operations are applied to a set, or when aggregating parts.

This paper explains fingerprint digital libraries for evidence and training in Section 2, integration and CO concepts using the 5S framework in Section 3, and a preliminary integration of the 4 DLs in Section 4.

2 The Different Fingerprint Digital Libraries

2.1 Recorded Prints

Recorded prints are the basis for the matching of images and the distorted images created by experiments. Large law enforcement databases may have millions of people's prints, where each one can come with 10 fingers, 10 toes, palm, pads

of feet, etc. There are direct and rolled prints, and sometimes repeated captures, including over time. The largest collections and systems generally are proprietary and not available to the public, or are related to scene analysis [5]. One of the biggest biometric database and fingerprint identification system is from the Federal Bureau of Investigation, at http://www.fbi.gov/hq/cjisd/iafis.htm. It has at least 66 million subjects in the criminal master file, along with more than 25 million civil print images. To determine whether two fingerprints match (Figure 2-A), examiners move beyond the common ridge patterns and focus on the unique and complex details of ridges that divide, cross, and terminate. This classification is based on four classes: terminations, bifurcations, trifurcations (or crossovers), and undetermined. The process of analysis and feature extraction from a single print can produce an enormous amount of information, like quality and direction maps, quality measures, etc. Besides the matching, there is assessment of image quality, e.g., based on NIST Fingerprint Image Quality (NFIQ) [6], considering details like direction, contrast, flow, and curvature.

2.2 Distorted or Synthetic Images

Distorted or synthetic images are created by algorithms that simulate motion and/or skin distortion. To investigate their effects on image quality, two types of distortions were considered: skin distortion [7] and blurring. The skin distortion model used in our initial experiments simulates skin plasticity around the contact point of a finger tip. It has 10 parameters controlling the model, as shown in Figure 2-B. The combination of a single recorded print with the 10 parameters, for example, can synthetically generate about 10,000 images. The blurring distortion model uses an increasing amount of blurring (Figure 2-C). The objective is to simulate several level of distortions, to compare levels of acceptable quality. Here one single image generates three other images.

2.3 Crime Scene

The evidence from a crime scene can come from thousands of people who visited a popular place, or touched an object, as shown in Figure 2-D, creating data which can be later compared with a criminal history record. Each person has ten fingers, and each finger can produce different images depending on the type of distortion. In addition, there are overlays of different prints, i.e., combinations of images from the fingers under the same substrate. The matching can process one fingerprint, multiple fingerprints, or combinations of entire and partial images against one database. Additional details can be present, regarding the fingerprint (location, orientation, size, pressure, distortion, etc.), the object touched (curvature, substrate, etc.), or methods of extraction and preservation. After the sample is collected, there is still the need to document the evidence history or provenance.

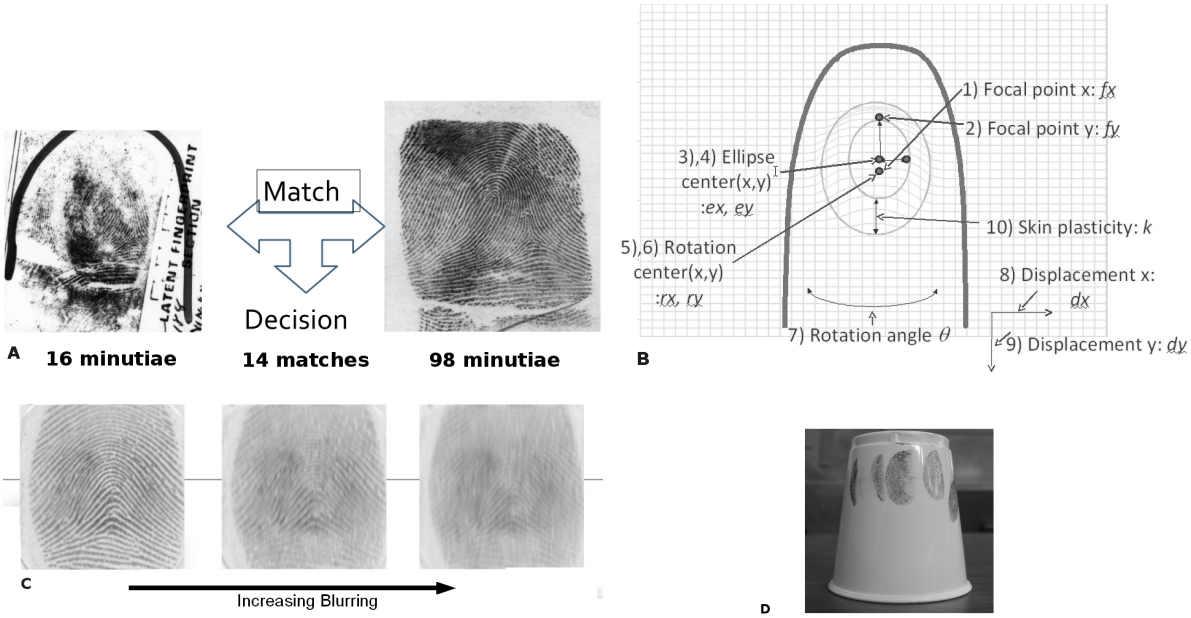


Fig. 2. (A) Matching Images. (B) Skin Distortion. (C) Blurring. (D) Crime Scene.

2.4 Training Materials

A first goal of the training effort is to develop templates for modules that encompass twenty different topics in biometrics, along with the use of combinations of examples to illustrate each of myriad types of situations. Ideally, for testing fingerprint examiners, the combination of examples identified could be used for assessment, so each case in an exam is distinct, reducing opportunities for cheating. The training modules will have examples for instruction, and yet others for exercises and examinations, taken from all of the other DLs. To give a sense of scale: Suppose that one image generates 100 distorted images. Multiply by 25 million possible suspects. Then try to match a crime scene image which has 55 partial fingerprints. Finally, select and link good examples for use in training.

We propose the use of compound objects, detailed in the next section, for connecting, aggregating, and re-using appropriate information in support of such large scale efforts.

3 Integration and Compound Objects in 5S Framework

3.1 COs and DL Integration

Agosti et al. [8] defined a Compound Object (CO) as a digital object that includes information about context, provenance, and relationships between resources. COs are aggregations of different information combined together in

order to shape a unique logical object. Several CO formats arise from different communities [9]. Even though there are a number of standards to support the management of COs, there is still incompatibility, motivating solutions for integration and interoperability.

Thus there is a second factor that needs to be analyzed: the process of integration. Kostas and Delis [10] divided the integration process into four steps: (i) discovery: systems "learn" about the existence of each other; (ii) identification: systems unambiguously identify their individual items; (iii) access: systems access their items; and (iv) utilization: systems synthesize their items.

In the case of COs, there is a fifth step, regarding how the objects are aggregated. The Dexter Hypertext Reference Model [11] for example, uses the "hidden structure approach", placing all of the data and all of the data interpretation inside the content portion of a component. The Amsterdam Model [12], on the other hand, uses the "separate structure approach", defining each piece of multimedia information as a separate block.

We propose to connect, reuse, and integrate COs, taking advantage of the 5S (Streams, Structures, Spaces, Scenarios, and Societies) framework, along with the 5S approach to integrate digital libraries; see http://si.dlib.vt.edu/. The integration of archaeological digital libraries has been described from the 5S perspective, but then we considered only digital objects, not their composition. Due to that work, we can build upon a well-documented and validated formal framework describing some of the essential aspects of digital libraries.

3.2 Concepts and Definitions

An **Integrated Digital Library** is a 4-tuple consisting of a union repository, a union catalog, union services, and a union society. The minimal union services of a digital library are represented by mapping and harvesting services, which are necessary to support integration. For the integration of COs, we can use the same definitions, considering the following aspects:

1. each CO has a handle, a structure, its internal components, and a boundary (so we clearly distinguish the CO from other objects) [3];
2. each CO has an interface or description which specifies how its information can be accessed;
3. each CO has a vocabulary to describe it and its internal composition;
4. the same vocabulary that is used to describe the components can be used for labels for the schema in the mapping service;
5. the mapping service is responsible for unique identification of all the objects in the union set;
6. the boundary is represented in the mapping service by what is within the internal structure of each CO;
7. the application should specify which approach is used for the object aggregation: "hidden structure approach", "separate structure approach", etc.

For the harvesting process, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) can be used, defining a mechanism for data providers

to expose their metadata. For disseminating the content in concert with a metadata harvesting protocol, some steps are necessary [13]: (i) wrap the data in a packaging format; (ii) include the metadata; (iii) encode the references to the files; and (iv) harvest the package. For this, OAI-ORE [14] or DCC [15] can be used, representing the objects and aggregations.

The complexity of the mapping and updating in the integration process can be affected by several factors, such as knowledge of the application domain, the number of elements in the local schema, and the size of the collection.

In the case of compound object technologies, such as DCC and OAI-ORE, the mapping process also depends on how the components are aggregated, what is their granularity, which vocabulary each technology is using, how the components are identified and structured, and how they are organized in a schema.

4 Integrating Fingerprint Digital Libraries

Building upon the fingerprint digital libraries summarized in section 2, and the compound object concepts described in section 3, this section presents the "discovery" and reuse/integration of the large amount of data present in fingerprint DLs.

We begin with an example of COs and the four initial sub-systems, as in Figure 3: (A) the recorded prints; (B) the distorted images; (C) the crime scene images; and (D) the training material, with suitable sequencing for pedagogy.

Compound Object 1 (CO1) has the following components: a fingerprint image from system A, one distorted image from system B, a crime scene image from system C, and a link to related training material, taken from system D. The components can be identified by CO1.A.1, CO1.B.1, CO1.C.1 and CO1.D.1, respectively. The CO1 structure can be represented by RDF, while the content could be packaged using OAI-ORE or DCC. The interface of CO1 can comprise the union information of its four components, along with the union of their respective vocabularies (individual, fingers, thumb, quality, distortion, parameters, etc.).

Further, DCC could be used to encapsulate the objects, or even OAI-ORE with a RDF parser, in an integrated DL service, providing *the match between latent and recorded fingerprints*, or *a chain of evidence to convince a jury of confidence of match*, for example. Other integrated DL services could consider *the object versions* (with the composition of distortions, for example) or *correspondence of versions with provenance*, in the crime scene application. Due to the amount of information and detail, these analyses would take longer if the services were not integrated.

Our preliminary results include: (i) an Entity-Relationship Diagram design; (ii) the implementation of the skin distortion model (Figure 2-B); (iii) testing of the blurring distortion (Figure 2-C); (iv) the description of internal steps of the NFIQ quality; and (v) an initial exploration of concepts that will be analyzed from the CO perspective. Though our project is in an early development stage, these preliminary results were important to highlight the amount of information

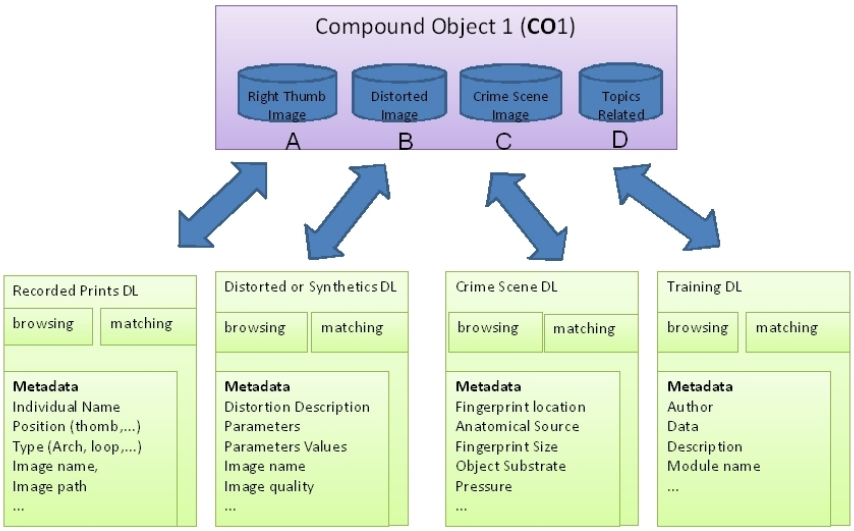


Fig. 3. An example of compound object using four digital libraries: (A) Recorded Prints, (B) Distorted Images, (C) Crime Scene Images, and (D) Training Material.

and details we need to manage, guiding us to explore the overall system by using a very large DL approach.

5 Summary and Conclusions

There are many integrations which relate to addressing large numbers of objects, considering combination, versions, and reuse of information. Our approach takes advantage of volume, concepts, and services already available and manageable in digital libraries.

We presented preliminary results for the integration of fingerprint digital libraries, along with an initial analysis from the compound object perspective. The following items were described: (i) four types of DLs (recorded prints, training materials, crime scenes, and experiments with distorted images); (ii) a summary description about the distortion models accomplished; (iii) examples of services available; and (iv) an initial analysis of CO integration concepts present in the 5S framework, along with minimum services such as harvesting and mapping.

Future work will further address the matching of latent vs. recorded prints, the determination of sufficiency and quality related to the matches, the analysis of other parameters/services for COs, and encapsulation and description using CO technologies.

6 Acknowledgments

We would like to thank CAPES (4479-09-2), FAPESP (2009/18438-7) and CNPq (481556/2009-5 and 306587/2009-2). We are grateful for grants from NIJ (National Institute of Justice) and BAE Systems. Thanks also go to NSF grants IIS-0910183, IIS-0916733, DUE-0840719, and CCF-0722259.

References

1. Jain, A.K., Maltoni, D.: Handbook of Fingerprint Recognition. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2003)
2. 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) 1–60
3. Murthy, U., Kozievitch, N.P., Leidig, J., Torres, R., Yang, S., Goncalves, M., Delcambre, L., Archer, D., Fox, E.A.: Extending the 5S Framework of Digital Libraries to support Complex Objects, Superimposed Information, and Content-Based Image Retrieval Services. Technical Report TR-10-05, Virginia Tech, Department of Computer Science (April 2010)
4. Schank, R.C.: Tell Me a Story: Narrative and Intelligence. Northwestern University Press (1995)
5. Gloe, T., Böhme, R.: The 'Dresden Image Database' for benchmarking digital image forensics. In: SAC '10: Proceedings of the 2010 ACM Symposium on Applied Computing, New York, NY, USA, ACM (2010) 1584–1590
6. Theofanos, M., Micheals, R., Scholtz, J., Morse, E., May, P.: Does habituation affect fingerprint quality? In: CHI '06 Extended abstracts on human factors in computing systems, New York, NY, USA, ACM (2006) 1427–1432
7. Maltoni, D., Cappelli, R.: Advances in fingerprint modeling. Image Vision Comput. **27**(3) (2009) 258–268
8. Agosti, M., Ferro, N., Silvello, G.: The design of a DLS for the management of very large collections of archival objects. First Workshop on Very Large Digital Libraries (2008)
9. Nelson, L., de Sompel, H.V.: IJDL special issue on complex digital objects: guest editors' introduction. International Journal of Digital Libraries **6**(2) (2006) 113–114
10. Saidis, K., Delis, A.: Integrating multi-dimensional information spaces. Second Workshop on Very Large Digital Libraries, Oct. (2009)
11. Grønb, K.: Composites in a Dexter-based hypermedia framework. In: ECHT '94: Proceedings of the 1994 ACM European conference on Hypermedia technology, New York, NY, USA, ACM (1994) 59–69
12. Gruzman, V.A., Senichkin, V.I.: Hypermedia models. Autom. Remote Control **62**(5) (2001) 677–694
13. Maslov, A., Mikeal, A., Phillips, S., Leggett, J., McFarland, M.: Adding OAI-ORE Support to Repository Platforms. Fourth International Conference on Open Repositories (2009)
14. Lagoze, C., Sompel, H.V.: Compound Information Objects: the OAI-ORE Perspective. Open Archives Initiative Object Reuse and Exchange, White Paper, http://www.openarchives.org/ore/documents (2007)
15. Santanchè, A., Medeiros, C.B.: A Component Model and Infrastructure for a Fluid Web. IEEE Transactions on Knowledge and Data Engineering **19**(2) (February 2007) 324–341

Capturing and Analyzing User Behavior in Large Digital Libraries

Giorgi Gvianishvili, Jean-Yves Le Meur, Tibor Šimko, Jérôme Caffaro, Ludmila Marian, Samuele Kaplun, Belinda Chan, and Martin Rajman

European Organization for Nuclear Research
CERN, IT Division,
CH-1211, Geneva 23, Switzerland
{giorgi.gvianishvili,jean-yves.le.meur,
tibor.simko,jerome.caffaro,
ludmila.marian,samuele.kaplun,belinda.chan}@cern.ch
http://www.cern.ch

Swiss Federal Institute of Technology
EPFL, Artificial Intelligence Laboratory,
CH-1015, Lausanne, Switzerland
martin.rajman@epfl.ch
http://www.epfl.ch

Abstract. The size of digital libraries is increasing, making navigation and access to information more challenging. Improving the system by observing the users' activities can help at providing better services to users of very large digital libraries. In this paper we explain how the Invenio open-source software, used by the CERN Document Server (CDS) allows fine grained logging of user behavior. In the first phase, the sequence of actions performed by users of CDS is captured, while in the second phase statistical data is calculated offline. This paper explains these two steps and the results. Although the analyzed system focuses on the high energy physics literature, the process could be applicable to other scientific communities, with and international, large user base.

Keywords: Invenio, CDS, Very large digital library, Log analysis, User behavior study analysis

1 Introduction

Digital libraries are playing a strategic role in the showcasing of research done by an institution or university, since 1988. Large scientific communities rely on the digital libraries as a primary resource for storing and acquiring information. One of the challenges is to make navigation in large amounts of data as intuitive as possible. Our goal is to concentrate on the users' specific needs in order to improve and optimize access to information. In addition to the successful survey done in 2008 on the information resources in High-Energy Physics [8] the behavior of CERN Document Server (CDS) users has been studied.

CDS is an instance of the Invenio software, which is developed and maintained at CERN. The number of records in CDS exceeds 1 million and continues to grow, while the number of unique users is more than 40 000, making CDS one of the largest digital libraries in the physics domain [14]. We have logged users' interaction with the search engine and analyzed them using automated data processing techniques. This automated approach helped to reveal important patterns, which are difficult or sometimes even impossible to spot by a human, due to the large amount of data involved.

In order to understand which options users can select and which of them were used or ignored, we first describe the CDS production environment, underling its core functionalities and possibilities. We then describe the logging phase and the type of information collected. Finally, we explain the automated extraction of additional information and the results obtainedreturned.

2 System Description

Invenio [1] [2] [3] [4] is a digital library system which is freely available under GNU General Public License. Invenio consists of a set of modules for maintaining intermediate to large digital library services. It has been actively used at CERN since 2002. Besides CERN, it is also used in diverse scientific institutions and universities worldwide like EPFL [5], DESY [6] and others. The system can handle not only articles and books, but also theses, photos, videos, etc.

Records maintained by Invenio are organized in collections that can be defined on top of any query. Users are offered either simple or advanced search interfaces. They can query specific fields, such as title, author, etc., sort the results or apply a ranking criteria (like word similarity).

Users can restrict their search to a set of specific collections or sub-collections, and the results returned can be merged into a single list.

Users can also customize the output format of the results: by default a summary of the results is displayed (brief HTML) but other formats such as detailed HTML, HTML MARC and others are also provided. Invenio has been translated into 26 languages and supports Unicode for information retrieval.

Users can also register an account in order to access restricted collections or to use Web 2.0-like services (baskets, alerts, etc.).

In CDS, there are approximately 8 000 registered users, representing a large portion of the high energy physics community. However, the majority of users are not registered ($\sim$ 40 000 users).

Most of the content maintained in CDS are articles and preprints, coming mainly from the high energy physics domain.

3 User Logging

3.1 First Phase

Invenio software allows user activities to be logged in real-time into MySQL tables. Standard logs collected from the web-server do not provide sufficient

information for observing users' interactions with the system. In addition to web-server logs we can store information about the query recall, the status of the user and the rank of downloaded documents. The main challenges arising in this phase are:

- Defining the data and the events to be captured
- Preserving the relationship between stored data
- Making logging transparent to end users

Two tables contain information about user queries, with the following data: user id, date, host name, IP address, HTTP referrer and query recall. Recall is stored as the list of records unique identifiers. Other tables are dedicated to log the downloads and the accesses to detailed page view information. For each record, download time, client host, user id, file format, HTTP referrer and display position are logged. The download table is used not only to store local file downloads, but also downloads of documents which are hosted on remote servers, and which cannot be extracted from the web-server logs. To preserve the relationship between stored data query id, user id and IP address are used. These identifiers can uniquely identify the history of a user's action in the system.

3.2 Second Phase

After the logs have been collected, a post-processing phase that is more computationally intensive is executed offline. Four types of counts are extracted from the logs:

- Number of detailed page views: for each record we count the occurrences of record abstract being viewed
- Number of downloads: for each record we count the occurrences of the associated full-text being downloaded
- Number of displays: for each record we count the occurrences of the record being listed on the results pages
- Number of seens: for each record we count the occurrences of record being seen. We mark all records seen from the first up to the one on which an action has been performed (download/view). For example, if a user downloads record #6 we mark all records from #1 up to #6 as seen, since those records would have most probably been seen by the user. This count provides us with an approximate result of records seen, since there is no guarantee that the user has really seen those records.

These numbers, can then be used not only for information but also for ranking and for analysis of the relationship among records. They might also suggest the reorganization of the digital library for optimizing its performance. Concerning ranking, combining these counts with other attributes like freshness, citation frequency and Hirsch index is being studied within the scope of the collaborative D-Rank [13] project at Swiss Federal Institute of Technology (EPFL) and Central European Organization for Nuclear Research (CERN). The core idea of the

project is to take into consideration a user's previous interaction with the system: if some subsets of documents have been downloaded or viewed, it is assumed that their importance will be preserved; on the other hand, documents which were displayed or potentially seen, but not downloaded or viewed will be considered as less important.

4 Analysis

After analyzing more than 130 000 queries maintained on CERN Document Server (CDS), it can be observed that 73.1% of users are using the English interface which is set by default. Usage of other languages is relatively equally distributed (Figure 1).

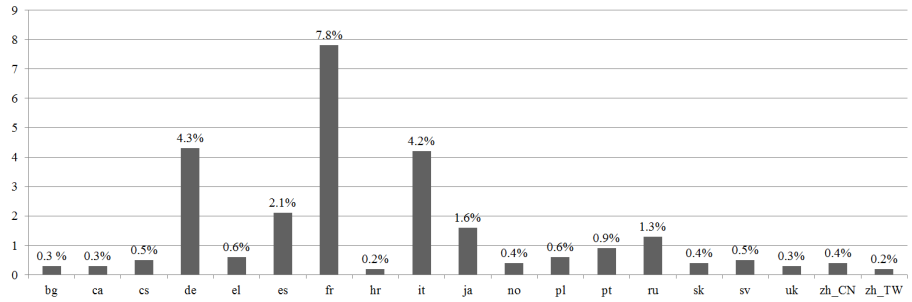


Fig. 1. Number of users using non-English search engine interface. The English interface has been used 48 507 times (73.1%).

The default ordering of the records is 'latest first', but it can be changed to display results according to word similarity criteria. As we observed, such type of adjustments are extremely rare: in the case of CDS only in 1% of the queries ranking method has been changed from latest first to word similarity. This confirms a habit that has already been observed in the past [9]. It confirms that professionals from a specific field look more for the recent publications [10]. Yet another explanation can be that options in the interface are not intuitive enough to users. Sorting has been used in less than 3% of the queries Table 1. The default setting where the descending order of results has been used is in 97% of cases.

CDS supports a wide variety of output formats, but users change the default setting (condensed display) in only 1.5% of all queries.

Advanced search was used in approximately 8% of all queries, using the default matching type most of the time (85.4%) (matching all the words). Remaining matching types and operators used are compared in Table 2 and Table 3. As we can observe from the Table 2, besides the default option users are most often using the 'Regular Expression' matching type. Such contrast of using

Table 1. Percentage of using various sorting criteria. Default ordering criteria (Latest First) was used in 97.2% of cases.

Latest First	Chronological Order	Key Title	Year	Report Number	Author	Title	Other
97.2%	1.0%	0.9%	0.3%	0.1%	0.1%	0.1%	0.3%

the default or the most advanced technique can be caused by using pre-defined queries. Operators used operators (Table 3) suggest that users prefer achieving higher precision by restricting the criteria (Google behavior). We can observe in Table 4 that the same fields are used to query CDS in bothe the simple and advanced search interfaces.

In Table 5 we can observe the 64 most often issued query terms. Although it is possible to enter Boolean expressions and years in the dedicated fields, users prefer typing them using free text query.

Table 2. Percentage of using various matching criteria, in advanced search.

Matching Type	Percentage
All of the words:	85.4
Any of the words:	1.3
Exact phrase:	1.5
Partial phrase:	0.3
Regular expression:	11.5

Table 3. Percentage of used operators in the advanced search for defining relationship among matching fields.

Operator	Percentage
AND	96.8
OR	2.3
NOT	0.9

Rank of downloads and detailed page views is shown in the Table 6. Top ranked records are downloaded/viewed on average 9 times more than ones on the 9th position. In Table 7 we can see the 10 most often displayed records with corresponding counts. The search engine returned no results in less than 1.5% of all queries. The distribution of user access through the day (Figure 2) confirms that CDS is the institutional repository.

Table 4. Percentage of using different fields in the simple and advanced search interfaces.

Field	Simple Search	Advanced Search
any field	49.1	71.8
author	15.2	12.8
title	19.5	6.5
keyword	2.8	3.5
report number	3.0	2.6
year	9.8	0.6
other	0.6	2.2

Table 5. List of most often used terms in user queries for 5 weeks, with corresponding frequencies. (Typically several terms are combined to form query.)

Term	Frequency	%	Term	Frequency	%	Term	Frequency	%
lhc	2289	3.5%	magnet	237	0.4%	energy	184	0.3%
cern	1468	2.2%	neutrino	235	0.4%	collision	184	0.3%
atlas	1324	2.0%	lhcb	234	0.4%	school	178	0.3%
physics	1023	1.5%	programme	233	0.4%	control	176	0.3%
higgs	469	0.7%	collaboration	220	0.3%	model	171	0.3%
particle	412	0.6%	trigger	218	0.3%	technical	171	0.3%
detector	404	0.6%	performance	213	0.3%	first	169	0.2%
alice	346	0.5%	quantum	213	0.3%	training	168	0.2%
data	337	0.5%	hadron	199	0.3%	electron	163	0.2%
beam	319	0.5%	bulletin	198	0.3%	tunnel	161	0.2%
lecture	299	0.5%	computing	194	0.3%	field	160	0.2%
design	287	0.4%	system	192	0.3%	experiment	158	0.2%
accelerator	284	0.4%	student	189	0.3%	academic	157	0.2%
muon	258	0.4%	collider	189	0.3%	logo	154	0.2%
theory	239	0.4%	introduction	187	0.3%	collisions	153	0.2%
calorimeter	237	0.4%	reconstruction	185	0.3%	john	147	0.2%

5 Conclusion

Thanks to its rich mechanisms Invenio is giving a lot of possibilities for observing how users are interacting with the system. The number of users and records maintained by CDS makes it one of the largest open access digital repository in science. Capturing and analyzing user logs can provide us with hints on how to improve the usability of the system. Log analysis results can be combined with other types of user behavior studies, for better understanding the user needs.

Log analysis procedure in CDS is done in two phases. In the first step logs are collected online. The second phase is run offline, for extracting detailed statistics.

Collected data can be applied to the new ranking algorithm, building recommendation systems or identifying user communities with common interests. Other possible applications are: construction of query expansion mechanisms, user interface optimization, identifying most requested queries for their opti-

Table 6. Rank on which records have been downloaded or detail page viewed, with corresponding counts. Mostly 'latest first' ordering has been used.

Rank of Results List	Download Frequency	Rank of Results List	Page View Frequency
1	1428	1	1885
2	566	2	973
3	353	3	768
4	287	4	618
5	203	5	494
6	180	6	359
7	143	7	381
8	128	8	261
9	117	9	297
$\geq$10	4175	$\geq$10	6676

Table 7. Top 10 most often displayed records, with corresponding seen, download and abstract view counts.

	Displays	Seens	Views	Downloads
1	237	17	22	198
2	247	23	10	130
3	358	54	24	100
4	182	9	10	97
5	139	4	5	80
6	154	9	36	76
7	238	25	26	75
8	234	17	15	63
9	106	6	0	63
10	76	4	2	58

mization, defining the most suitable time for running computationally intensive tasks and many others.

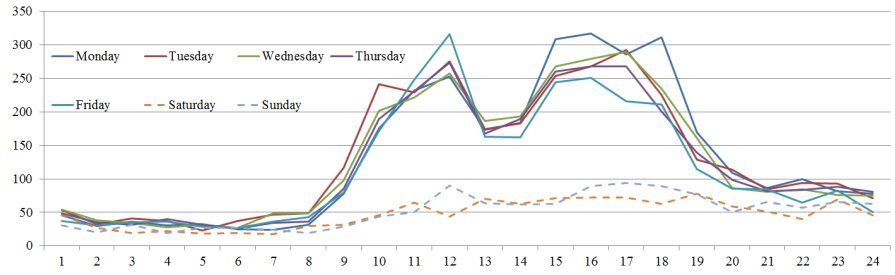


Fig. 2. Average user access through the day. By dashed lines there are denoted weekends, while with the solid line workdays.

References

1. Invenio software website, `http://invenio-software.org`
2. CERN Document Server, `http://cds.cern.ch`
3. Pepe A., Le Meur J.-Y., Simko. T.: Dissemination of scientific results in High Energy Physics: the CERN Document Server vision, (2006)
4. Pepe A., Baron T., Gracco M., Le Meur J.-Y., Robinson N., Simko T., Vesely M.: CERN Document Server Software: the integrated digital library, (2005)
5. Swiss Federal Institute of Technology, Knoweledge and Information Services `http://infoscience.epfl.ch/`
6. DESY, A Research Center of the Helmholtz Association `http://desy.de/`
7. SLAC, Stanford Linear Accelerator Center Library/SPIRES, Stanford University `http://www.slac.stanford.edu/spires/`
8. Gentil-Beccot A., Mele S., Holtkamp A., O'Connell H. B., Brooks T.C.: Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course, (2008)
9. Jones T., Cunningham S.J., Mcnab R., Boddie S. : A Transaction Log Analysis of a Digital Library , Department of Computer Science, University of Waikato, International Journal on Digital Libraries, pp. 152–169 (1999)
10. Tenopir C.: Use and users of electronic library resources: An overview and analysis of recent research studies. Washington, DC: Council on Library and Information Resources, (2003)
11. Covey D. T.: Usage and usability assessment: Library practices and concerns. Washington, DC: Council on Library and Information Resources, (2002)
12. Papatheodorou C., Kapidaki S., Sfakakis M., Vassiliou A. : Mining User Communities in Digital Libraries, (2003)
13. Vesely M., Rajman M., Le Meur J.-Y., Using Bibliographic Knowledge for Ranking in Scientific Publication Databases, Published in IOS Press, (2008)
14. Ranking Web of Worl Repositories `http://repositories.webometrics.info/top800_rep_inst.asp`

The Kramerius System – Open Source Solution for Digital Libraries

Tomas Foltyn

National Library of the Czech Republic

Tomas.Foltyn@nkp.cz

Abstract. The Kramerius System and homonymous project belong to the most important activities of the National Library of the Czech Republic. In the present time more than 7 500 000 pages are available via this digital library. What is unique in the broader European context is the fact that Kramerius is used by almost 30 libraries in the Czech Republic and one installation is available also in Slovakia. This enables effective way of the data replication and cooperation in system development.

Keywords: digital library, open-source, digitization, metadata

1 Introduction

The main objective of this short article is to describe the way of presentation of digital documents in the National Library as well in the biggest libraries in the Czech Republic. It describes past development of the Kramerius System version 3, which is now really wide-spread in the Czech Republic, and the current development of the version 4, which is based on Fedora Commons. The first release is going to make public in the late August with the idea of the successive enlargement not only in Czech Republic territory.

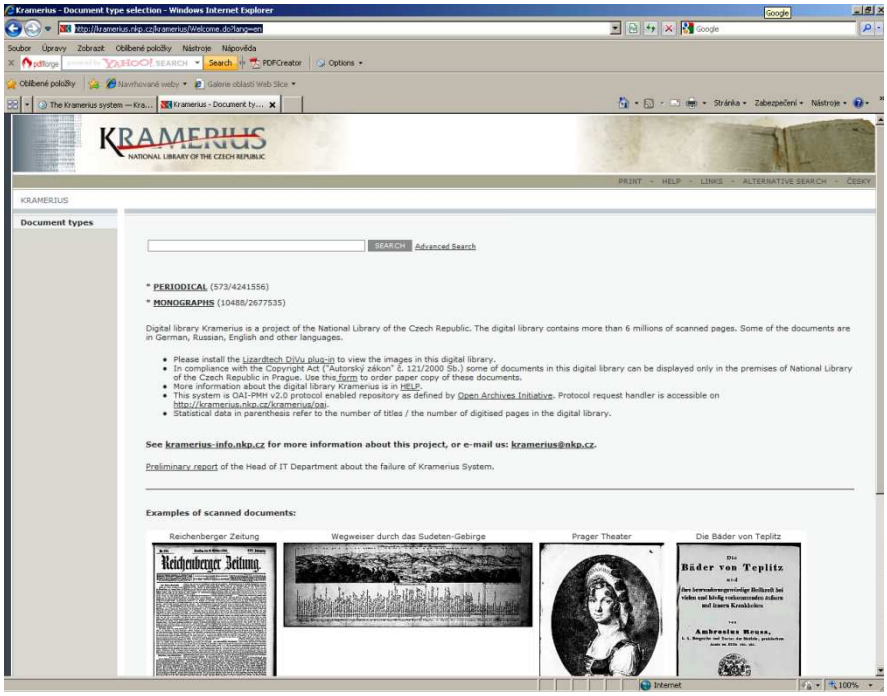
2 Kramerius System background

Besides the Manuscriptorium[1] and WEBarchive[2] Kramerius[3] is the third huge digitization project of the National Library of the Czech Republic. Primarily it is aimed to the digitization of modern books and periodicals, which are endangered by the degradation of the paper, secondarily to the other documents generating the national cultural heritage. The Kramerius System is a special Content Management

[1] See the official Manuscriptorium webpage http://beta.manuscriptorium.com/.
[2] See the official Englisch web page http://en.webarchiv.cz/
[3] Please see the webpage http://kramerius.nkp.cz/.

System (CMS)[4], which is intended for making digitized material accessible. It can serve either in local library network or on the whole internet. The system was developed as an open source application on the basis of the GNU GPL license, so it is freely available to every institution, which would like to have its own digital library.[5] The first impulse for the Kramerius system creation were the enormous floods in the Czech Republic in 2002 and a large number of destroyed or damaged books. The reformating was used as a mean of the salvage or replacement of damaged documents, and it was necessary to create a tool that would make these digital copies accessible en masse. Physical media (CD-Rs or DVDs) were not considered as a suitable solution.

Fig. 1. The front page of the Kramerius System English version in the National Library of the Czech Republic[6]



The Kramerius System was developed continuously in the close cooperation with the Academy of Sciences Library[7] and private company Qbizm technologies[8] and

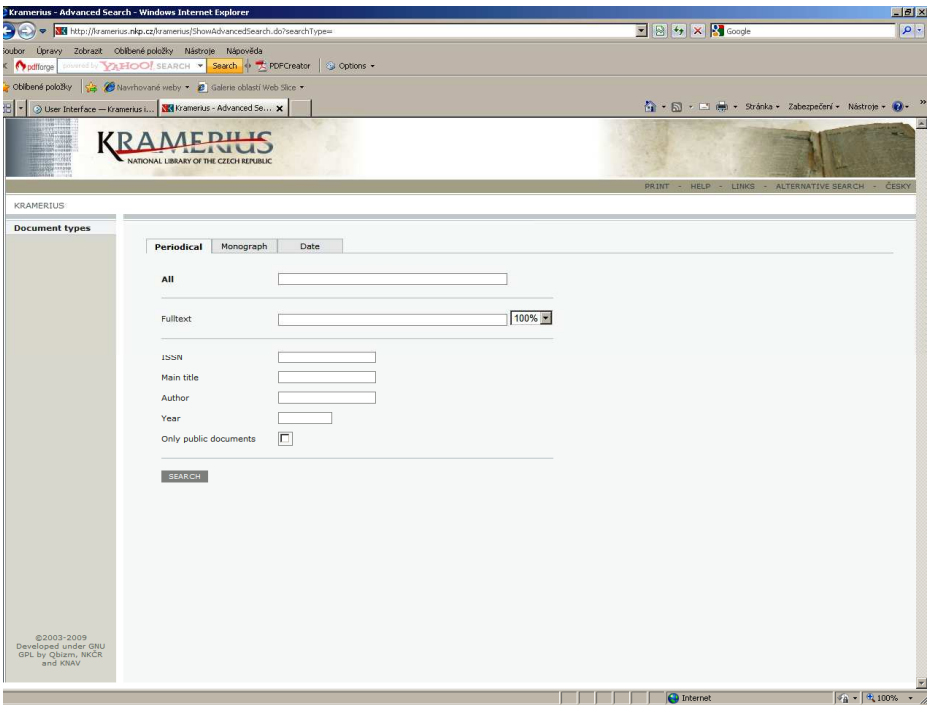
[4]The definition of CMS is available for example on the website http://en.wikipedia.org/wiki/Content_management_system.
[5] All the technical requirements are available on this webpage http://kramerius.qbizm.cz/menu/Podpora/FAQ.html.
[6] See http://kramerius.nkp.cz/kramerius/Welcome.do?lang=en.
[7] See http://www.lib.cas.cz/en.
[8] See http://www.qbizm.com/.

founded from various number of resources. The last version is called Kramerius 3.3.1 and is available both in Czech and English version. According to the users requests (based on the user questionnaire published in August 2009) the last development of the third version has begun in the end of the last year. This latest news is concerned in the modification of search engine (e. g. full-text searching through selected title, issue etc.), alphabetical order of specific letters (e. g. ä, ü, ö) and especially in the RSS feeds creation, which is based on MDT classification.

3 Interfaces, technical description and file formats

The Kramerius System has two separate interfaces – one public aimed to the common user access, and second one for the professional work of system administrators. The user interface allows to the final users browse via digital library. The users have more possibilities how to do it – they could select the special types of document – periodicals or monographs – and afterwards browse through the list of authors or titles to the requested document.

Fig. 2. The advanced search field of the Kramerius System[9]



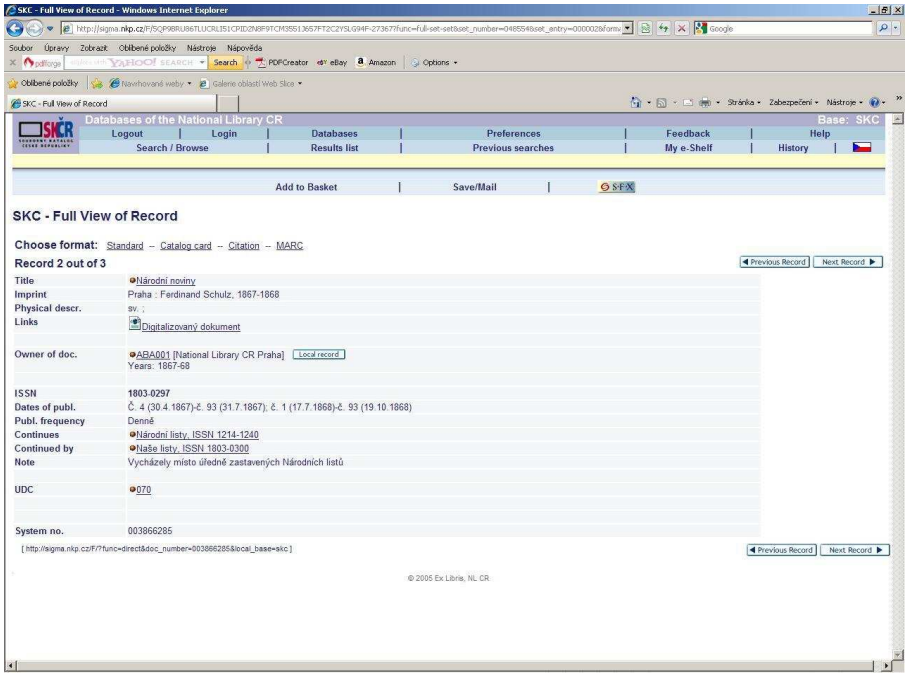
Other possibility is to use the integrated special search engine Lucene,[10] where each user could write required keyword and let the system search the term either in full text

[9] See http://kramerius.nkp.cz/kramerius/Welcome.do?lang=en.

or after filling the term into right box at the requested level. The date searching is also possible.

The administrative interface serves to the administrators of the system for the everyday "real" work with the system. It contains many important operation fields, which are important for daily use. Administrators could import and export the data in various number of ways (e. g. only images, images and metadata etc.), grant the access rights to other users, make the documents accessible or on the other hand hide them, start the statistics or replicate the documents to the other institutions using the Kramerius System. That is very important – via replication the libraries could share the already digitized data electronically (of course in accordance with the copyright). The system is also equipped with the OAI PMH protocol for harvesting large amounts of metadata and text files by cooperating institutions or international activities.

Fig. 3. Print-screen from Union Catalogue, where is possible to see the link to digitized document.[11]



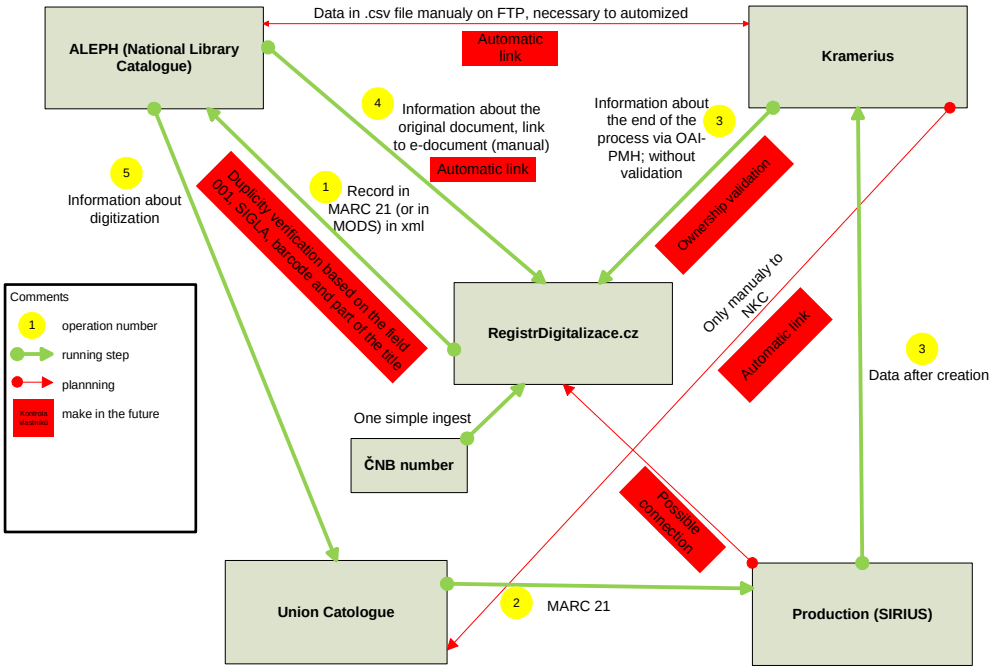
The great advantage of the Kramerius System is also the connection with the catalogue system Aleph and the digitization registry.[12] This connection works on both

[10] For technical details see http://lucene.apache.org/java/docs/.

[11] See e. g. http://aleph.nkp.cz/F/VV9Y5X425P2MGACDPVVT6AVGSPMKISH8T4QJALCX4U24GM2A8M-25771?func=full-set-set&set_number=012234&set_entry=000097&format=999&CON_LNG=ENG.

sides. Final user could search document either in Union Catalogue or National Library Catalogue and after one click see the digitized document, or open the relevant record in the catalogue directly from the digitized document. At the present time this connection functions fully in the case of periodicals. The connection between Kramerius and Digitization Registry[13] works also very well.

Fig. 4. Schema of the library systems connection[14]



Each document in the Kramerius System is published in DJVu format, which was chosen according its capability to the size reduction. But there are also some disadvantages – the biggest one is that DJVu requires that plug-in need to be installed. Users could obtain also limited PDF file. As metadata standard using for import is used the internal DTD standard for monographs and periodicals, which is based on UNIMARC, but it is not the full UNIMARC.[15] In the Kramerius System there are also available OCR results in simple txt format, which are used for the full-text searching. They could be also exported to METS. A document in this format consists of several sections (header, bibliographic metadata, administrative metadata, files, structural maps, structural links etc.). The bibliographic metadata are exported in the MARC XML and DUBLIN CORE formats, whereas the administrative metadata are exported

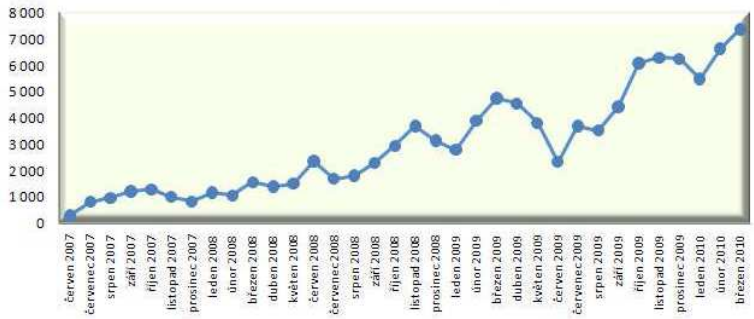
[12] For detailed information about this project please see next chapter.
[13] See the Digitization Registry webpage www.registrdigitalizace.cz.
[14] See http://kramerius.nkp.cz/kramerius/PShowVolume.do?id=5126&it=0.
[15] DTD documentation is available on the website http://digit.nkp.cz/techstandards_cz.html.

in the PREMIS and MIX formats. For descriptive, structural and administrative metadata xml format is used.

4 Numbers

Now, there are more than 7,5 million pages available in the Kramerius System of the National Library of the Czech Republic – about 4 630 000 pages of periodicals and 2 925 000 pages of monographs.[16] Every user could browse via 614 periodical titles and more than 11 000 volumes of monographs. From this huge number of pages it is about 5 845 000 pages OCRed. All the documents are accessible according to the copyright. From the users point of view and from the information from statistics it is known that there is approximately 7 000 unique accesses and almost 1 500 000 hits per month.

Fig. 5. Diagram of the number of unique visitors per month[17]



5 Kramerius – national program and information webpage

Kramerius is not only the name of the digital library, but also the alternative name of the national cooperation research program VISK 7, which is funded by the Ministry of Culture of the Czech Republic.From this funding many titles were digitized and made accessible. In the VISK 7 research program cooperates about 25 Czech libraries that use mostly the Kramerius System as their digital library and thus it is easy to replicate data. The National Library stores their data and is responsible for their preservation.

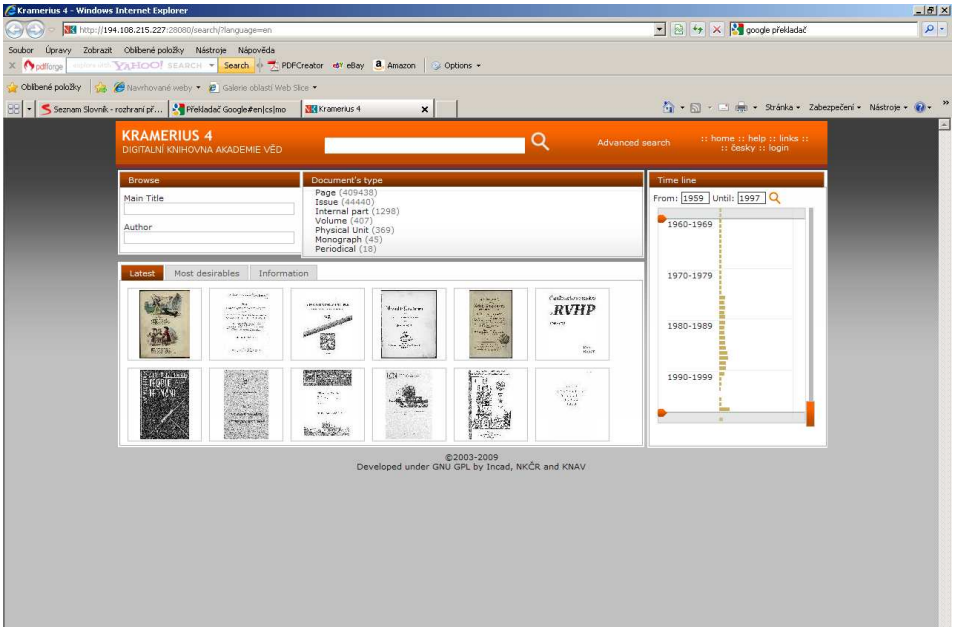
For improvement of user awareness the Kramerius information portal was created at the end of 2008.[18] Here everybody is able to find the information about the

[16] Numbers to the end of June 2010.
[17] See the context in the presentation of Dr. Polisensky available http://osliusi.mlp.cz/doku.php/program.
[18] Official webpage is http://kramerius-info.nkp.cz/.

digitization of the modern collections of the National Library of the Czech Republic, national and international projects, topical information on the documents that have been made accessible, events and conferences or about the standards used in other libraries. That could be helpful also for professional public. This portal is available both in Czech and English version.

6 New version of Kramerius System development

Fig. 6. Front page of the Kramerius System version 4. Academy of Sciences Library localization.



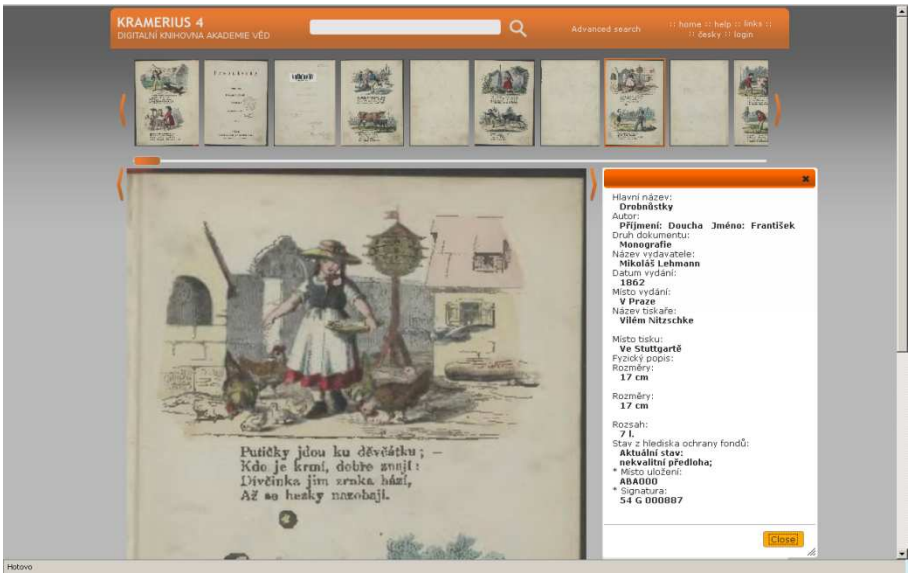
Current development of Kramerius system in version 4 is aimed to the utilization with FEDORA commons, which is established under the close cooperation of the National Library of Czech Republic, Academy of Sciences Library and private company INCAD.[19] This version should be for example more friendly for final users, should have more possibilities for user rights establishment and should be able to work with more formats, but, and it is very important, by keeping the existing functionality. This continuity with the existing system was from the beginning of the development an essential principle, which is respected till now.

The new system, which is now fully tested in both above mentioned libraries, is based on the JAVA JDK 1.6 – so it is possible to use it as web application in any J2EE container (e. g. Apache Tomcat). In the second version the web portal solution as

[19] See http://www.incad.cz.

Liferay should be included. As search engines SOLR and LUCENE are used. Web interface was created via JQuery and GWT. Data are managed through UUID identification and internal Fedora data streams.

Fig. 7. The Kramerius System version 4 title "Drobnustky" image and basic metadata description. Academy of Sciences Library localization



The version 4 is as well as the previous versions of the Kramerius System available as open-source solution. It is freely for disposal both as the software solution (resource code included) or virtual image (VMware or V-box solution). Kramerius in version 4 is more opened to larger amount of file formats (not only DJVu, but also PDFs, JPEGs etc.) and also other types of documents (not only periodicals and monographs, but also maps, manuscripts, graphics etc.). It will be finalized in two languages mutation – Czech and English. Developers of the system will continue immediately after first public release on additional upgrades – at the first level e.g. on the implementation of ALTOxml text files or advance administrative menu.

7 Conclusion

From all the above mentioned details it is clear, that this new solution will be applicable in all the libraries, which are looking for good sophisticated solution, that is capable to work with large amount of digital data and it is user friendly. There are no limits to extend the Kramerius system worldwide.

Error Tolerant Large Scale FRBRization

Andreas Juffinger[1], Elisabeth Lex[2], and Nuno Freire[3]

[1] The European Library, PO Box 90407
2509 LK The Hague, Netherlands
andreas.juffinger@kb.nl,

[2] Know-Center GmbH, Inffeldgasse 21a,
8010 Graz, Austria
elex@know-center.at

[3] Biblioteca Nacional, Campo Grande 83,
1749-081 Lisboa, Portugal
nfreire@bnportugal.pt

Abstract. Digital Library Discovery Interfaces Online Public Access Catalog, and Web Search Engines do have nowadays one common purpose: to provide a simple, easy to use interface for users to find relevant material in the underlying information space. Users are most often interested in finding a distinct intellectual work in the first place and subsequently getting information about different variants such as expressions and manifestations. However, library cataloging has always focused on producing meaningful holding information for archiving and retrieval purposes of single objects like books, journals or maps. The problem of efficiently mapping multiple catalogue records to a single intellectual work is very complex. On an data aggregator level, where one wants to combine records from different institutions with different cataloguing rules, this becomes even more challenging. Comparing each record with all other records leads to a quadratic complexity and is for large scale digital libraries unfeasible. In this paper we present a graph based solution for error tolerant grouping of bibliographic records into work-sets. The method minimizes human efforts and is able to deal with different cataloguing practices and common mistakes. Additionally, we present a locality-sensitive hashing method for bibliographic records to minimize the number of necessary complex comparisons allowing the method to scale nearly linearly.

1 Introduction

The European Library is a free service that offers access to the resources and catalogues of the 49 national libraries of the Council of Europe in 35 languages. Resources can be both digital object (books, posters, maps, sound recordings, videos, etc.) and bibliographical data. Recently The European Library formed a strategic partnership with LIBER and CERL to provide researchers from the humanities with all necessary content and catalogue information. The mission of

The European Library is to provide access to all bibliographic records of these libraries. One major challenge in aggregating the 49 national libraries and about 500 research libraries, is the identification of duplicates and the clustering of similar work. This de-duplication is crucial to provide a useful discovery interface otherwise a search result would consist of endless lists of duplicates.

Functional Requirements for Bibliographic Resources (FRBR) is the most promising attempt in the library domain to tackle nowadays user needs for bibliographic data. In the FRBR study[9] the functional requirements for bibliographic records are defined in relation to generic user tasks when searching and making use of library catalogues. FRBR as a conceptual model is intended to be independent of any cataloging code or implementation [12]. The FRBR model defines three different group of entities: a) Group 1: entities and primary relationships whereby the entities work, expression, manifestation and items represent products of intellectual or artistic endeavour. b) Group 2: entities and responsibilities relationships with entities like persons and corporate bodies in addition to group 1 entities, and c) Group 3: entities and subject relationships adding subjects to group one and two. The group 1 entities are further defined as work (distinct intellectual or artistic work - the story behind a book), expression (the specific form a work takes each time it is realized - the realization of a story as a textual work), manifestation (the physical embodiment of and expression - the hard cover printed textual expression), and item (a single instance of the hard cover manifestation). The aim of this paper is de duplication in an discovery information system, therefore we focus on the work instance, with the overall goal to propose only distinct works to the end users in a search result list. Figure 1 shows the appropriate entities and relations of group one and two. Note that this paper focuses on primary and responsibility relationships only, and leaves the entity subject relationships to further work.

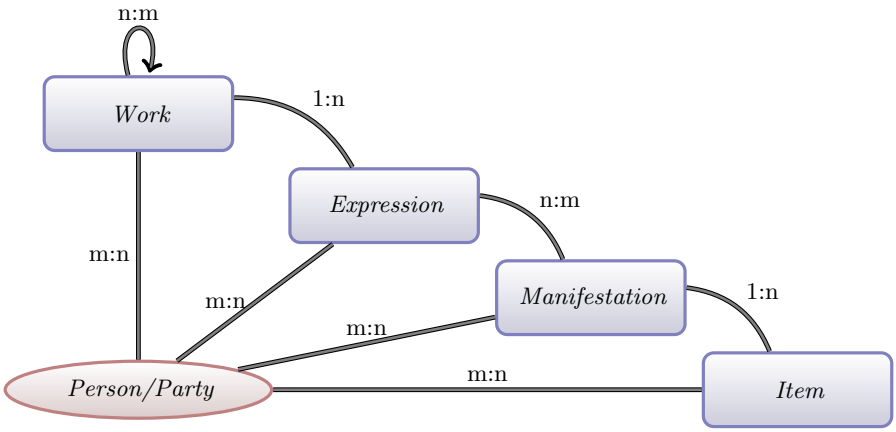


Fig. 1. FRBR conceptual model: entities, relations and responsibilities

For discovery interfaces it is crucial to provide users the possibility to browse the bibliographic resources starting at a single work or original source. Cataloguing in libraries always focused on the description of the two bottom FRBR entities: manifestations and items. Therefore, in a library database, each bibliographic record describes a manifestation, and may have associated several items (when the library hold several copies of the same book). Expression and work data also exist in library databases, but these where recorded only as attributes of the manifestations, not as entities. If a book has several editions, a bibliographic record is created for each edition, but no relationships will exist between these records because expression and work are just represented as attributes of the manifestations. We face now two sides of the same coin, firstly, the cataloging side where different manifestations and items are described without a reference to a unique intellectual work, and secondly, the user side where most people are not particularly interested in a specific expression or manifestation of an item. FRBRization of bibliographic records is meant to fill this gap and aims to group bibliographic records into so called FRBR work sets. As outlined in Hickey and Toves[6], a work set is defined as the set of all bibliographic records of the same intellectual work. A general procedure to retrieve work-sets is given by the following algorithm:

Input: Set of bibliographic records R, list $P = []$, result set $W = \{\}$
Output: $W = \{w_1, w_2, ..., w_m\}$ work-sets $\{\forall r_a, r_b \in w_i \Rightarrow r_a = r_b\}$
for $r \in R$ **do**
  $r' = preprocess(r)$;
  $P \leftarrow P \cup r'$;
end
$sort(P)$; $w = \{P[0]\}$;
for $r' \in P$ **do**
  if $compare(w[0], r') == 0$ **then**
    $w \leftarrow w \cup r'$
  else
    $W \leftarrow W \cup w$
    $w = \{r'\}$;
  end
end

Algorithm 1: Work-Set Algorithm

In and ideal world (assuming *preprocess*() and *compare*() are constant time) this problem is straight forward and computable with a worst case time complexity of $\Theta(n \log n)$ comparisons ruled by the complexity of comparison based sorting. Hickey and Toves[6] algorithm runs on real world scenarios with 54 Mill. bibliographic records in reasonable time. The authors describe their procedure for MARC21 bibliographic records, what does not allow the identification of work-sets across heterogeneous bibliographic formats without additional work on the algorithm. Another drawback of this methodology is that the computation of the keys is rather complex, due to the number of hardcoded rules and the involvement of authority lookups.

In this paper we present the model and the appropriate algorithmic framework to compute work sets in highly heterogenous aggregator environments. The main requirements for aggregators are:

1. Support for heterogenous and proprietary data formats, cataloguing rules and pronunciations of bibliographic records. This implies the identification of near duplicates and clusters of bibliographic records rather than identical sets.
2. Scaleability to billions of bibliographic records. At this scale, it necessary to guarantee that preprocessing steps and comparisons are constant time and cheap.

Note that string matching of authors and cosine similarity calculation without further optimizations are generally linear in the length of the text or number of tokens and are expensive operations at this scale. (see Section 3)

Although nearly every library holds different expressions and manifestations of the same work the problem of duplication becomes significantly more complex in our setting when dealing with national libraries. Whereby public libraries usually do not hold many different versions of the same work, the mandate of most national libraries is to collect everything published in and about a country. Depending on collection strategies the national libraries therefore collect reprints, editions and different variations of the same work. When it comes to the level of combining content of different national libraries one degree of duplication happens whenever a book about another country (B) is published in country (A), then at least two national libraries (A+B) collect the same item. From the perspective of FRBR work, a translation into another language is still the same work, to combine these records its then necessary to map different attributes, for instance title and original title, or even to combine different pronunciations of authors for instance (Tolstoy, Aleksey Nikolayevich and Tolstoj, Aleksej Nikolajevic) and, at the encoding level, different character sets.

The rest of this paper is structured as follows: In Section 2 we introduce our graph based model for bibliographic records and the error tolerant similarity, as a solution for requirement (1). Section 4 introduces locality sensitive hashing and the selected class of hash functions to ensure the scalability requirement (2). The conclusion and further work is then presented in Section 5.

2 A Graph Model for Bibliographic Records

Nearly all national and research libraries follow one of the following cataloguing rules which are based on, or similar to, the International Standard Bibliographic Description (ISBD). The ISBD rules define bibliographic descriptions in eight areas, whereby the most important area for this work are areas one (title and statement of responsibility area) and two (the edition area). On top of the ISBD rules, the most commonly used cataloguing rules are Anglo-American Cataloguing Rules (AACR2) with its successor Resource Description and Access (RDA), and Rules for Alphabetical Cataloging (RAK). Unfortunately, this fast number

of rules and specifications still leave space for variations in their interpretations and on a practical level their complexity often leads to misunderstandings and error proneness. To highlight the complexity of such rules an example: ISBD defines a number of punctuations which allow a cataloger to enter additional information in the title field of a bibliographic record. One such rule is: "The general material designation is enclosed in square brackets, the first bracket being preceded and the second followed by a space (␣[ ]␣)." IFLA provides a mapping of ISBD elements to FRBR entities and attributes[4]. This ISBD mapping has manifestations as the primarily FRBR target. This becomes clear due to the fact that libraries catalogues reflect real resources, therefore the manifestations of a FRBR work. But FRBR is defined in top down and unfortunately, it does not explicitly state how one can derive a work from a manifestation. Remember that FRBR work is defined as a distinct intellectual work of a person or group. The most common practice is therefore to identify a work by its title and authors (see Work-Set Algorithm[5] or Variations3 Algorithm[6]).

2.1 Graph Construction

The variations in cataloging and data formats make it necessary to define a model for bibliographic records which can easily be adapted to different data formats and cataloging practices allowing some degree of error tolerance. To our best knowledge there is only one appropriate data structure with the necessary expressiveness: graphs. Graphs are the most general data structure in computer and information science. A graph models relations among elements of a set. Elements can be related to each other in every possible way. Unrelated elements can be seen as unconnected nodes in a graph. A sequence of elements or a list can be seen as very simple graph. Trees or other hierarchical arrangements are special cases of graphs. Formally a graph $G(V, E, \Psi)$ consists of a finite set of nodes, or vertices V, a finite set of edges, or relations E, whereby $\Psi : E \mapsto \{X \subseteq V : |X| = d, d \geq 2\}$. Graphs with $d = 2$, so called 2-graphs are the most common and mathematically best studied form. Note that the rest of this paper exclusively deals with 2-graphs, or graphs for short.

The following bibliographic records (see Figure 2) are taken from our database and serve as an example of how we create the bibliopgraphic graphs, which are then the basis for our calculations. The examples also shows that only title/author remain the same throughout different manifestations.

Clearly the records (1)-(3) are the same intellectual work by J.R.R. Tolkien and (4) is a book about the author. For a human it is not to hard to identify the two separate intellectual works and the relations between each other. But grouping (1)-(3) as the same work is not trivial for computers. The authors are not the same for all three records (record 2 has different authors), and the titles vary in the order and number of terms. A common practice is to apply a number

[4] June 2010: http://www.ifla.org/files/cataloguing/isbd/isbd-frbr-mapping.pdf
[5] June 2010: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm
[6] June 2010: http://wiki.dlib.indiana.edu/confluence/x/OQBGBQ

```
(1) Tolkien, J. R. R. (1955). The Fellowship of the ring. Lord of
the Rings. part 1. NY: Ballantine.
(2) Tolkien, John R. R., Lee, A. (2008). Lord of the rings. part
1. The fellowship of the ring. London: HarperCollins.
(3) Tolkien, J. R. R. (2008). La comunidad del anillo = The
fellowship of the ring / Lord of the rings, part 1. Barcelona:
Minotauro.
(4) Grotta, D. (1976). J.R.R. Tolkien: Architect of Middle Earth :
a biography. Philadelphia: Running Press.
```

Fig. 2. Examples of bibliographic records as graph

of normalization steps to make the titles similar and then use a boolean similarity measure to check for similarity. But such normalization leads to a huge number of rules, hard to maintain and understand with unpredictable side effects and valid just for a single dataset or cataloguing institution.

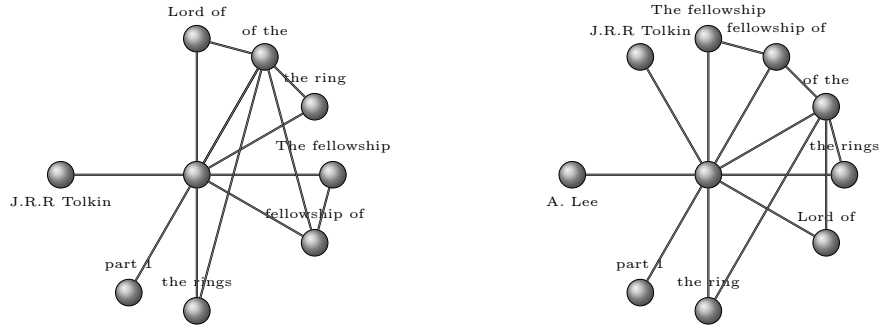
Instead of making the records similar we propose the use of an error tolerant distance measure based on the unmodified bibliographic record. We then measure the dissimilarity and then tune the threshold for optimal results. Our initial implementation of this idea was based on the cosine similarity between bibliographic records, but it lead to results which have been significantly worse than the rule based method. The two main reasons for this are, on one hand that the bag of words approach completely ignores the structure of the bibliographic records, and on the other hand the amount of text in bibliographic records is very small. Therefore probabilistic methods do not perform at their best.

We therefore implemented a graph based representation, where the structure remains, but no proprietary rules are necessary. The graph construction algorithm to transform a bibliographic record into a graph for FRBR work implies the following steps:

1. Split the bibliographic record into parts according to ISBD rules. (by .␣).
2. Create shingles of size 2 within each group.
3. Connect each shingle with the artificial center node.
4. Connect consecutive shingles.
5. Merge identical shingles.

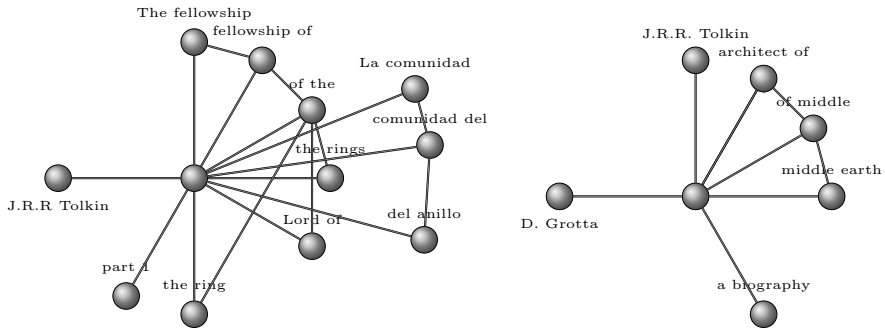
Examples for such bibliographic graphs are shown in Fig. 3 and Fig. 4. Fig. 3(a) shows the graph for the bibliographic record (1). The algorithm splits the original record into four parts: (a) The author: J.R.R. Tolkien, (b) the first part of the title (c) the second part of the title, and (d) the part information. Each part is then tokenized by white spaces and all shingles of size two are created. For instance the first title part becomes the set of shingles {{the fellowship},{fellowship of}, {of the}, {the ring}}. Each shingle then gets connected with the root node (center of graph). Furthermore we draw connections from each shingle in an ordered set to its successor. In a last step we ensure unique shingles by merging identical shingles.

Fig. 4(a) shows a more complex example: a translation of the book into a different language. The original title is also provided in the bibliographic record,



(a) Graph: J.R.R.Tolkin (1) (b) Graph: J.R.R.Tolkin (2)

Fig. 3. Examples of two editions of the same work



(a) Graph: J.R.R.Tolkin (3) (b) Graph: Tolkin Biography

Fig. 4. Examples of translation and biography

so we are able to construct a graph containing information on both titles. From these examples one can see that the graphs for record (1)-(3) share lots of things in common, but are not isomorph.

2.2 Graph Edit Distance

The graph edit distance is formally introduced in Tsai and Fu [13] as inexact graph matching. Note that inexact graph matching is the superordinate concept of graph edit distance and includes all methods which deal with distances between graphs. Tsai and Fu [13] formulated this problem and gave a widely used definition for inexact graph matching. Their paper focused on structurally isomorphic graphs and deals therefore only with node and edge substitution. A method which also deals with node and edge deletions and insertions, respectively, has been introduced by the authors in [14]. Wong et al. [15] improved their method in 1990 by taking also future costs into account. In Sanifelu and Fu [11] the authors introduced the idea of graph edit distance in the context of inexact graph matching. The graph edit distance, also known as error correcting distance, gives a numerical estimator for the similarity or dissimilarity between two graphs, and is nowadays widely accepted as a similarity measure for graphs. The graph edit distance is defined as the minimum number of edit operations required to transform one graph into the other. The graph edit distance further assigns each of the operations: insertion, deletion, and substitution of nodes and edges, a certain cost. The cost is usually inversely proportional to the likelihood of a certain edit operation. Naturally, the cost function has an important influence on the similarity of graphs. Furthermore, in [2], Bunke shows that the graph edit distance is equivalent to graph isomorphism under certain cost functions. A direct result of this is the NP completeness of the general graph edit distance with such cost functions, because graph isomorphism is known to be NP hard.

This is the reason that the graph edit distance is not widely used. A common misunderstanding is although that this worst case performance applies to all graphs. Dickinson et al. [4] have been able to proof, that for graphs with unique node labels the edit distance has quadratic time complexity. Graphs with unique node labels are thereby defined as all graphs $G(V, E)$ which satisfy the following condition: For any pair $u, v \in V$ of labelled nodes $\alpha(u)$ and $\alpha(v)$ these graphs satisfy $\alpha(u) = \alpha(y) \Rightarrow u = v$ and $\alpha(u) \neq \alpha(y) \Rightarrow u \neq v$. Furthermore, the authors assume that the underlying alphabet of node labels is an ordered set, for example, the integers $\{1, 2, 3, ...\}$. Note that our aforementioned algorithm to create graphs for bibliographic records ensures unique node labels in step (5).

2.3 Bibliographic Distance

The first step in calculating the graph edit distance is the alignment of existing nodes. Although it would be possible to integrate this into the graph edit distance calculation it self we have introduced a preprocessing step where we align existing nodes. In this step we calculate the maximal matching with minimal costs c_{AB} between the nodes of the two graphs whereby the cost function states

the similarity. The Hungarian Method by H.W. Kuhn [8] computes the maximum weight matching in a bipartite graph. Discussion of the algorithm details are too lengthy to present here; for a summary of the algorithm see A. Frank [5].

The costs are calculated according to the normalized distance function $f_d(n, m)$ and the Levenstein distance $d(n, m)$

$$f_d(n, m) = \frac{d(n, m)}{\min(|n|, |m|)} \tag{1}$$

After the calculation of the optimal matching we filter out node matchings which are to dissimilar (≥ 0.20). At the end of this process we now have two graphs, whereby we know which nodes to map and how expensive this mapping is c'_{AB}.

The second step is then to calculate the graph edit distance between two bibliographic graphs. The distance is defined as the sum of costs for node and edge insertions and deletions. We have defined unique costs of 1 for the different operations normalized by the appropriate size ($n_{insert} = n_{delete} = \min(\frac{1}{|V_A|}, \frac{1}{|V_B|})$, $e_{insert} = e_{delete} = \min(\frac{1}{|E_A|}, \frac{1}{|E_B|})$). The total edit distance is then calculated as the sum of all operations:

$$e_{AB} = \sum n_{insert} + \sum n_{delete} + \sum e_{insert} + \sum e_{delete} \tag{2}$$

The graph based bibliographic edit distance s_{AB} between two records A and B is then defined as the weighted sum of the costs for filtered node matching c'_{AB} plus the graph edit distance between the bibliographic graphs e_{AB}.

$$s_{AB} = \beta c'_{AB} + (1 - \beta) e_{AB} \tag{3}$$

Utilizing this bibliographic distance we have been able to achieve significant better performance for the identification of work-sets as with the rule based methods across different datasets. Additionally, we have been able to minimize the necessary effort of domain experts for constructing rules and complex "if-then-else" statements, solving requirement (1).

This method is rather complex and comparing two bibliographic records with each other is of $O(k^3)$ with an average number of k shingles per record. It is not valid to assume cheap constant time for comparison, so requirement (2) is still unsolved. After a more in-depth complexity analysis in the next section we therefore introduce a bounding box method which allows cheap and constant time comparison (see Section 4). Whenever two records are within the same bounding box, then we perform the complex comparison based on bibliographic edit distance. This two level approach leads asymptotically to a constant time complexity for comparing records.

3 Complexity Analysis and Algorithm Comparison

The work-set algorithm (see Algorithm 1) can operate basically on three different levels:

1. string level, operating on a character level, whereby the *preprocess* method produces a string representation of the record and the *compare* method compares these representations.
2. token level, whereby the overall method operates in the vector space model dealing with tokens. Preprocess leads therefore to a set or ordered set of tokens as representation of the bibliographic record.
3. concept level, whereby the bibliographic record is transformed into a semantic space and comparison happens then on the level of comparing concepts.

The algorithm of Hickey and Toves[6] falls into the level one. The calculation of the author/title keys is, after authority normalization, rule based and constant time c_k. Comparing two strings is linear time in the length of the involved strings. Assuming that the length of the title field as well as the number and length of author names is limited, we can safely assume that comparing two keys is constant time c_c. The computation of work-sets for the whole dataset of size n takes then $c_k\Theta(n) + c_c\Theta(n \log n) + c_c\Theta(n)$.

Our graph based method is a level two method. ISBD splitting, tokenization, shingling and building the graphs is linear time in the number of tokens. With the same argument as above, the title length and author number is limited, the graph construction becomes constant time c_g. Comparing two graphs is cubic in the number of tokens, but again, for a limited number of tokens this is constant time c_e. This leads to the same complexity as above with different constants $c_g\Theta(n) + c_e\Theta(n \log n) + c_e\Theta(n)$.

Given the same asymptotical complexity it comes down to comparing the constants. Looking deeper into c_k reveals, that there are a number of complex string manipulations stripping certain parts, removing things in brackets (ISBD like) and also according to the Name Authority Cooperative Program NACO (http://www.loc.gov/catdir/pcc/naco/normrule.html). All these steps depend on operations on the whole string - so we end up with $O(l * m)$, whereby l is the number of times the whole string needs to be visited and m is the length of the string. Our graph based method includes splitting according to ISBD rules and tokenization on a string level which is less complex. The creation of the graph is then linear in the number of shingles s. Therefore the overall complexity of preparing the graph than doing rule based normalization is significantly smaller $m/d * s * t \simeq c_g << c_k \simeq l * m$. Assuming an average token length of $d = 10$ characters per token, $s = 8$ shingles per record and $t = 4$ linear shingle operations. On the other hand we assume only $l = 10$ complex string operations what leads to $640 \simeq c_g < c_k \simeq 2000$.

Given similar costs for comparison $c_c \simeq m = 200$ and $c_e \simeq s^3 = 192$ the constants are basically in the same range but both algorithms do not hold for requirement (2). String comparison and edit distance involve simply to many operations in comparison to integer comparison. Given the huge number of necessary comparisons we need a way to identify when its worth to apply these expensive operations. Locality Sensitive Hashing (LSH) is one way to achieve this. LSH can be used to estimate the necessity of a complex comparison by one step integer comparisons.

4 Locality Sensitive Hashing

The key idea behind locality-sensitive hashing (LSH) is to hash the records using several hash functions to ensure that for each function the probability of collision is much higher for objects that are close to each other than for those that are far apart. Then, one can determine near neighbors by hashing the query point and retrieving elements stored in buckets containing that point.

4.1 LSH and the nearby neighbor problem

In the context of LSH constructing work-sets is defined as a nearby neighbor search within the vicinity of a bibliographic record. Each record within a work-set is closer to all records within the same work-set as to any other record. From an algorithmic point of view this problem is the decision version of the nearest neighbor problem, a well known optimization problem. More formally p is a R-near neighbor of a point q if the distance between the two points is at most R (for details see [1, 7]).

As outlined in Adoni and Indyk [1], the LSH algorithm relies on the existence of locality-sensitive hash functions. Let $\mathcal{H}$ be a family of hash functions mapping d to some universe U. For any two points p and q, consider a process in which we choose a function h from $\mathcal{H}$ uniformly at random, and analyze the probability that $h(p) = h(q)$. The family $\mathcal{H}$ is called locality sensitive (with proper parameters) if it satisfies the following condition: (a) if $\|p-q\| \leq R$ then $Pr_{\mathcal{H}}[h(q) = h(p)] \geq P_1$ and (b) if $\|p-q\| \geq cR$ then $Pr_{\mathcal{H}}[h(q) = h(p)] \leq P_2$. In order for a locality-sensitive hash (LSH) family $\mathcal{H}$ to be useful, it has further to satisfy $P_1 > P_2$. However, one typically cannot use $\mathcal{H}$ as is since the gap between the probabilities P_1 and P_2 could be quite small. Given a family $\mathcal{H}$ of hash functions the authors in [1] recommend to amplify the gap between the high probability P_1 and low probability P_2 by concatenating several functions. In particular, for parameters k and L they choose L functions $g_j(q) = (h_{1,j}(q), \ldots, h_{k,j}(q))$, where $h_{t,j}(1 \leq t \leq k, 1 \leq j \leq L)$ are chosen independently and uniformly at random from $\mathcal{H}$. These are the actual functions that are use to hash the data points.

For text related problems Charikar [3] has defined the following LSH family $\mathcal{H}_c$: Pick a random unit-length vector $u \in \mathcal{R}^d$ and define $h_u(p) = sign(u \cdot p)$. Such a hash function can also be viewed as partitioning the space into two half-spaces by a randomly chosen hyperplane. The proofed theorem states that the probability that a random hyperplane separates two vectors is directly proportional to the angle between the two vectors. The probability that a randomly chosen hyperplane is separates two vectors is much higher when the points are far apart, and much smaller when the points are closed to each other. Ravichandran et al.[10] further have shown the following implicit form:

$$\cos(\theta(u, v)) = \cos\left(\pi * (1 - Pr[h_r(u) = h_r(v)])\right) \tag{4}$$

This equation states an alternate method for finding cosine similarity. As we generate more number of random vectors, we can estimate the cosine similarity

between two vectors more accurately. Ravichandran et al. further outline the relation of this to the hamming distance, which can be calculated very fast and easy.

4.2 LSH as bounding box for work-sets

The LSH family $\mathcal{H}_b$ for bibliographic data is based on the definition of Charikar [3] and Ravichandran et al.[10]. Our vector space is the unigram term space without stop word removal and stemming. Further we take only author and title parts into account and ignoring other parts in the bibliographic records. Note that due to the nature of bibliographic records, which consists of author names and terms from multiple languages, the term space is extremely large. But records are very short, therefore we need to select a relatively high unit-length d for our random vectors. Additionally, we define the LSH family $\mathcal{H}_b$ as:

$$\mathcal{H}_b = \mathcal{H}_a \cup \mathcal{H}_t \cup \mathcal{H}_c \tag{5}$$

Whereby the $\mathcal{H}_a$ family consists of unit vectors which are generated on the author name vector space only, the $\mathcal{H}_t$ family consists of the unit vectors generated based on the title parts, and the $\mathcal{H}_c$ family is based on both parts. This separation also allows us to define different parameters for each subspace.

Although we are still experimenting with different parameter sets, we have been able to reduce the number of necessary complex edit distance calculations per bibliographic record by a factor > 1000 with high recall (guaranteeing that we do not overlook candidates). Furthermore, we store the hash values in a database what makes it technically very easy and fast to search the vicinity of a bibliographic record. With respect to computational complexity LSH guarantees a constant number of constant time comparisons, what clearly solves the requirement (2).

5 Conclusion and Further Work

In this paper we have outlined and argued that the use of an error tolerant measure to build FRBR work sets has significant advantages over extensive normalization and error intolerant comparison. The main points are less human interaction necessary, easily extendable to new datasets, and of course tolerant to different spellings and typos. Our graph based method does not ignore the structure of bibliographic records and therefore outperforms simple bag of words approaches. On the other side our method calculates the edit distance which allows variations in the pronunciations and the structure and is therefore more tolerant as rule based methods. Our graph based bibliographic edit distance lies therefore in-between unstructured and completely structured methods.

With LSH we are able to limit the number of necessary complex comparisons. The described method is applicable also for rule based and bag of words comparisons and guarantees a minimum constant time comparison of records reducing thereby significantly the overall computational costs.

To evaluate and compare our bibliographic edit distance method meaningful with other methods we are currently building a test corpus with the help of domain experts. This test corpus should then allow to tune parameters but also to compare different strategies.

References

1. A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Foundations of Computer Science*, 51(1):117, January 2006.
2. H. Bunke. Error Correcting Graph Matching: On the influence of the underlying cost function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):917–922, 1999.
3. M. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, page 388, New York, New York, USA, 2002. ACM.
4. P. Dickinson, H. Bunke, A. Dadej, and M. Kraetzl. On graphs with unique node labels. *Proceedings of IAPR Workshop*, pages 13–23, 2003.
5. A. Frank. On Kuhn's Hungarian Method - A tribute from Hungary, 2004.
6. T. Hickey and J. Toves. FRBR Work-Set Algorithm. *Dublin: OCLC*, 2003.
7. P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM*, pages 604–613, New York, New York, USA, 1998. ACM Press.
8. H. Kuhn. The Hungarian Method for the assigment problem. *Naval Research Logistic*, pages 83–97, 1955.
9. O. Madison, J. Byrym, S. Jouguelet, D. McGarry, N. Williamson, M. Witt, T. Delsey, E. Dulabahn, E. Svenonius, and B. Tillet. Functional Requirements for Bibliographic Records, January 2009.
10. D. Ravichandran, P. Pantel, and E. Hovy. Randomized algorithms and NLP: Using Locality Sensitive Hash Functions for High Speed Noun Clustering. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 629, Morristown, NJ, USA, 2005. Association for Computational Linguistics.
11. A. Sanfeliu and K. Fu. A Distance Measure between Attributed Relational Graphs for Pattern Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 1983.
12. B. Tillett. A conceptual model for the Bibliographic Universe, 2003.
13. W. Tsai and K. Fu. Error-correcting isomorphisms of attributed relational graphs for pattern analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 75(12):757–768, 1979.
14. W. Tsai and K. Fu. Subgraph error-correcting isomorphism of attributed for syntactic pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 1983.
15. A. Wong, M. You, and S. Chan. An algorithm for graph optimal monomorphism. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):628–638, 1990.