Konferenzbeiträge | Atti | Proceedings

# Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009

## AT4DL 2009

**8th September 2009**
**Trento, Italy**

Edited by
Raffaella Bernardi, Sally Chambers and Björn Gottfried

Bozen · Bolzano University Press

# PROCEEDINGS OF THE WORKSHOP ON ADVANCED TECHNOLOGIES FOR DIGITAL LIBRARIES 2009

# AT4DL 2009

**8th September 2009**
**Trento, Italy**

Edited by
RAFFAELLA BERNARDI, SALLY CHAMBERS AND BJÖRN GOTTFRIED

BOZEN · BOLZANO UNIVERSITY PRESS

Editors
Raffaella Bernardi, Sally Chambers, Björn Gottfried

Cover design
Gruppe Gut Gestaltung, Bozen/Bolzano

Universitätsbibliothek Bozen
Biblioteca Universitaria di Bolzano
University Library of Bozen/Bolzano
Bozen/Bolzano University Press
Universitätsplatz 1 / Piazza Università 1
I-39100 Bozen/Bolzano
T: +39 0471 012 300
F: +39 0471 012 309
http://www.unibz.it/universitypress
universitypress@unibz.it

ISBN 978-88-6046-030-1

Available at:
http://purl.org/bzup/publications/9788860460301



# Preface

This volume contains the papers presented at AT4DL 2009 (*Workshop on Advanced Technologies for Digital Libraries 2009*) held on the 8th September in Trento (Italy) as a satellite event of ICSD (the *International Conference on Semantic web and Digital libraries.*)

The aim of the workshop was to bring together stakeholders from the digital library community in order to present an overview of state-of-the-art systems in the field and identify open research problems that require further work.

We thank the invited speaker, Andy Powell, for having accepted our invitation to give a talk on *what current Web trends tell us about digital library services* with a special emphasis on open, social and linked data. A special thanks goes to all the Programme Committee members who have been extremely efficient in their task, always punctual in submitting their contributions and accurate with their reviews.

Furthermore, we would like to thank the ICSD Conference Chair, Paolo Bouquet, for his enthusiastic response to our proposal of organising the workshop as satellite event of ICSD, the local organiser chair, Heiko Stoermer, for his prompt, helpful and clear response to all our questions and the University of Trento for hosting us.

The organisation of the workshop has been simplified enormously thanks to the EasyChair system which made the submission procedure, the review process and the editing of the proceedings smooth.

Our final thanks goes to The European Commissions eContent*plus* Programme for funding the projects in the Digital Libraries theme that make the state-of-the-art developments presented at AT4DL possible; to the CACAO project and the Free University of Bozen-Bolzano for undertaking a key coordination role in this event and to the Bozen-Bolzano University Press, based in the University Library, for designing and producing these proceedings.

August 2009

Raffaella Bernardi
Sally Chambers
Björn Gottfried

## Programme Chairs

Raffaella Bernardi
Sally Chambers
Björn Gottfried

## Programme Committee

José Borbinha
Vittore Casarosa
Carl Demeyere
Stefan Gradmann
Jakub Heller
Antoine Isaac
Udo Kruschwitz
Patrice Landry
Andreas Lattner
Mikolaj Leszczuk
Bernardo Magnini
Stefan Pletschacher
Massimo Poesio
Pasquale Savino
Viliam Simko
Massimo Zancanaro
Maarten de Rijke

# Table of Contents

# CACAO System: An Overview

Raffaella Bernardi[2], Massimo Balestrieri[1], Alessio Bosca[3], Luca Dini[3], Daniele Gobbetti[2], and Frédérique Segond[4]

[1] Gonetwork s.r.l.,
[2] Faculty of Computer Science, Free University of Bozen-Bolzano,
[3] CELI, s.r.l.,
[4] Xerox Research Centre Europe

**Abstract.** This extended abstract gives an overview of the CACAO system highlighting its innovative features, its current status and its next expected steps. CACAO is a two year project funded by the eContent*plus* Program carried out in collaboration by experts in Language Technologies and Digital Libraries.

## 1 Background

The large amount of on-line catalogues and digitalized documents across Europe launches two main challenges tightly connected to each other. On the one hand, there is the need of allowing users to search through the different catalogues simultaneously and on the other hand to find books on relevant topics even if in different languages. Language Technologies (LT) help achieving both tasks. An interesting example of how they can improve searchability in aggregated collections is provided in [5]; the system we present in this paper, CACAO[5] (Cross-language Access to Catalogues And On-line libraries), is mainly an answer to the multilingual retrieval challenge, though it necessarily concerns the aggregation problem too. Both tasks require a close collaboration between LT and (Digital) Libraries experts. CACAO is the result of joint efforts by experts of both fields.

CACAO project proposes an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries and library catalogues. By coupling sound Natural Language Processing techniques with available information retrieval systems the project aims at the delivery of a non-intrusive infrastructure to be integrated with current OPAC and digital libraries. The result of such integration will be the possibility for the user to type in queries in his/her own language and retrieve volumes and documents in any available language.

The system has been already put at work over the catalogues of the five consortium libraries[6] and tested on The European Library catalogues.

[5] CACAO is an EU project supported by the eContent*plus* Programme of the European Commission (ECP 2006 DILI 510035): `http://www.cacaoproject.eu`

[6] The Göttingen University Library (German), Library of the Free University of Bozen-Bolzano (Italian, English and German), Cité des Sciences et de l'Industrie (French), Kornik Library (Polish), National Széchényi Library (Hungarian).

The main objective of CACAO was crossing the chasm between sound innovation and adoption by library institutions for real life purposes. The test cases mentioned above show the success of the undertaken approach and the achievement of CACAO main goal. The system can now be further improved and fine-tuned thanks to the built infrastructure, its application to several different catalogues, and the collection of end-users queries' logs that we will be able to obtain through the time.

## 2 The CACAO solution

The architecture of the CACAO system, summarized in Figure 1, is an integration of several subsystems coordinated by a central manager that triggers scheduled activities (i.e. data harvesting or processing) and reacts to external stimuli represented by end users queries. The "Harvesting" subsystem is in charge of collecting data from digital libraries, abstracting from the multiplicity of standards and protocols, and storing them in a repository. The "Corpus Analysis" subsystem performs specific analysis and transformations on the data collected from libraries and infers new information that is then used to support query processing and resource retrieval (e.g. query expansion, terms disambiguation). The "CLIR" (Cross Language Information Retrieval) subsystem is in charge of analyzing the monolingual user query in input and transforming and enriching it by means of translations and expansions. Finally, the "Web Services" subsystem represents external modules providing specific services (e.g. linguistic analysis, translations).

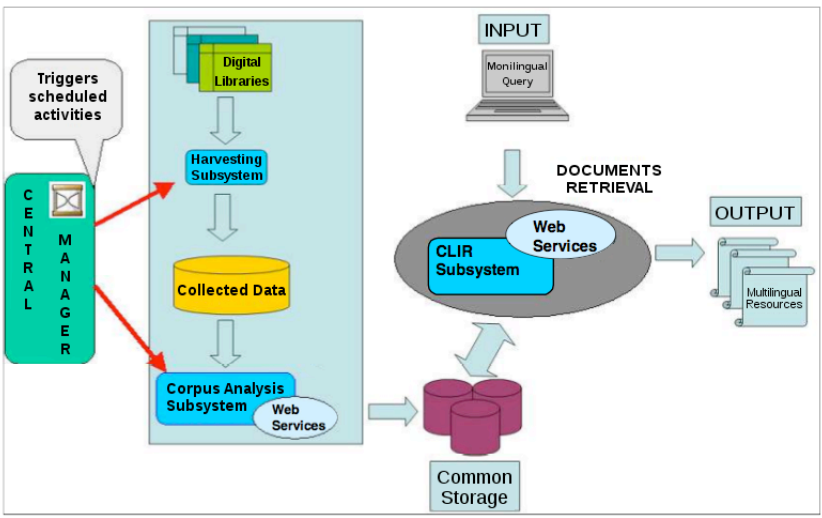



**Fig. 1.** CACAO architecture

The CLIR subsystem currently handles the languages of the consortium libraries, viz. German, Italian, English, French, Polish and Hungarian, and its flexible architecture allows easy integration of new resources in order to support new languages too.

In order to monitor its development CACAO has been evaluated already in its early stage via the participation to the TEL@CLEF campaign in 2008 [2]; the dataset of this campaign has been further used to evaluate the role of the different modules and resources that were developed during the project period. In particular, it has been used to enhance the system with the TLike algorithm, a translation algorithm that, by identifying in users logs whether a query is a likely translation of a previously submitted one, enriches CACAO translation resources. (See [3] for details.) A second module which is under evaluation for the different CACAO languages is the "Word To Category" (W2C) module. It has been described in [1] where a first evaluation in a controlled experiment has been reported for German, Italian and English.

CACAO can either be installed locally to obtain a Cross Language Access to the Catalogue of a certain library, or used in a federated setting. In both cases, users can access it via a Google like simple interface provided by the consortium and easy customizable, or via a facet-based advanced interface. Both interfaces will be soon available via the project web site `http://www.cacaoproject.eu`. Currently, the consortium is completing the harvest of TEL records and constructing three Thematic Portals on European History, Geography and Mathematics. To facilitate these aggregations, CACAO has developed an Application Profile based on The European Library Application Profile for Objects, discussed in [4].

## 3 Conclusion

CACAO experience shows that LT are in a mature stage for being applied to real life tasks, as the one required by the Digital Libraries world; furthermore such tasks and the possibility to analyse real users' behaviours and requests launch interesting new challenges and increase the appeal of these research field. CACAO is already running over five libraries catalogues covering six European languages, and is currently harvesting data from TEL. Therefore, it will soon become a useful source for log analysis and any further evaluation of LT related modules that can be easily integrated into its architecture. The consortium is now at work on collecting end-users feedback, as well as evaluating the TLike algorithm and the W2C module that are expected to further improve CACAO precision.

## References

1. R. Bernardi, D. Gobbetti, and L. Siciliano. Multilingual access to library catalogues: Word sense disambiguation via classification systems. In *Proceedings of the International Conference on Semantic Web and Digital Libraries*, 2009. In Printing.
2. A. Bosca and L. Dini. Query expansion via library classification systems. In *CLEF@TEL*, 2008.
3. A. Bosca and L. Dini. The role of logs in improving cross language access in digital libaries. In *Proceedings of the International Conference on Semantic Web and Digital Libraries*, 2009. In Printing.
4. B. Levergood, L. Siciliano, and S. Chambers. An application profile for interoperable and reusable metadata in a cross-language context. In R. Bernardi, S. Chambers, and B. Gottfried, editors, *Proceedings of The Workshop on Advanced Technologies for Digital Libraries (AT4DL 2009)*. University Library of Bozen/Bolzano, 2009. This volume.
5. D. Newman, K. Hagedorn, C. Chemudugunta, and P. Smyth. Subject metadata enrichment using statistical topic models. In *JCDL*, 2007.

# Digital Aeschylus
# Breadth and Depth Issues in Digital Libraries

Federico Boschetti

CIMeC, University of Trento, Italy
federico.boschetti@unitn.it

**Abstract.** Digital Libraries can grow along two different dimensions: breadth and depth. In the first case, works of many authors extend the existing collections. In the second case, different editions of the same works and related studies populate a monothematic region of the library. The *Digital Aeschylus Project* is aimed to collect, link and process digital objects based on primary and secondary sources related to the ancient Greek tragic poet.

## 1 Digital Libraries and Philological Needs

The most complete Greek and Latin corpora of texts, such as the Thesaurus Linguae Grecae (TLG) and the Packard Humanities Institute (PHI) Latin collection, are based on authoritative, most recent critical editions of each classical author. In these collections, only the text established by the editor is digitized, whereas the critical apparatus is omitted. Such approach to the ancient text, just about acceptable for literary and linguistic purposes, is unfeasable for philological studies. In fact, the philologist needs to identify manuscript variants and scholars' conjectures, in order to evaluate which is the most probable textual reading, accepting or rejecting the hypotheses of the previous editors. Furthermore, he or she needs to examine the commentaries, articles and monographs concerning specific parts of the text. Thus, the extension in breadth of the aforementioned collections needs to be integrated by the extension in depth, according to the paradigms of a new generation of digital libraries (see [7] and [18]).

In order to go in depth, philological studies are necessarily focused on single authors, genres or periods, even if they need to find links and parallels in the entire Greek and Latin literature. For this reason, teams of specialists need to share a common infrastructure, as pointed out by [8]).

For instance, the [15] Project is building a cyberinfrastructure to interrelate different philological and archeological projects. The [13] Project, on the other hand, has created a large platform to manage textual variants of Latin texts. The [12] Project, even if it is focused on a single ancient author, has developed a suite of services that can be easily extended to other authors.

The *Digital Aeschylus* Project, <http://www.himeros.eu/digitalaeschylus>, aims to provide philologists with a search engine on primary and secondary sources for the study of the Athenian tragic poet's tradition, taking into account

the standards for annotation and textual references that are emerging in the domain of the digital philology.

## 2 Structure of the Digital Aeschylus Project

The project is structured in modules, concerning the bibliographical catalogation, the acquisition of digital images of the documents, manual transcription and annotation of the most relevant manuscripts and early printed editions, OCR of recent editions, information extraction from the digitized documents. Due to the loose interdependence of the modules, they can be easily developed asynchronously.

### 2.1 Bibliographical Catalogue

The bibliographical catalogue related to manuscripts, printed editions and studies on Aeschylus aims to extend and integrate the printed repertory edited by [20]. In particular, it will supply the references to the digital resources on Aeschylus provided by large and general purpose digital libraries, such as [9] and [10], or provided directly by the research team, directed by V. Citti, that is working to the new edition of Aeschylus' tragedies.

The catalogue can be considered the roadmap for the digitization process, because it registers which resources are online and which ones are not yet accessible.

### 2.2 Digital Diplomatic Editions

Digital images of Aeschylean manuscripts and early editions have been collected under the direction of V. Citti and M. Taufer (see [19], for the current status of the acquisition).

The second step concerns the transcription and annotation of the most relevant materials, according to the Text Encoding Initiative guidelines related to manuscripts and early editions, [6]. These digital documents can be automatically collated, in order to identify textual variants and collect them in dynamic critical apparatus. (Techniques for automated collation by multiple alignment algorithms are illustrated in [17]). Furthermore, the different layouts of the manuscripts can be compared, going along with the recent interest for the colometric assessment of the tragic choral parts. In fact, it is demonstrated that the study of the colometry, i. e. the disposition of the strophes in different lines, sheds light on the generation of transmission errors.

Unfortunately, as demonstrated in [16], the optical character recognition on early editions is still unsatisfactory. For this reason, editions printed before the XIX century must be manually transcribed like the manuscripts. Naturally, the trascription is not performed by scratch, but by modification of a digital copy of a recent edition.

### 2.3 OCR on Recent Printed Editions

OCR can be applied to XIX and XX century critical editions, reaching up to 99% of accuracy on the text and more than 90% of accuracy on the critical apparatus. Anyway, it is important to point out that the critical apparatus is approximately only 5% of the page in editions with minimal information, and approximately 14% of the page for more informative editions.

These performances are obtained by the alignment and merging of three different OCR outputs and the application of an automated system of spell-checking, supported by the evidence of the OCR outputs. After suitable training, both [1] FineReader 9.0 and [14] 0.3 are able to recognize polytonic Greek characters mixed to Latin characters, whereas [2] 4.1 is able to recognize only polytonic Greek. Each OCR engine is more or less reliable for specific characters, and the reliability is evaluated by training sets. The merging system computes the most probable character in each position: the result significatively overwhelms the performances of the single engines.

The details of the system are illustrated in [5] and a similar approach is exposed in [11].

### 2.4 Information Extraction from the Repertories of Conjectures

Repertories of conjectures register not only the corrections to the ancient text suggested by the editors in their own editions, but also the proposals for emendation contained in commentaries and articles. As illustrated in [3], the repertories of conjectures have a trivial structure: in fact, more than the 90% of the items are constitued by the reference to the verse affected, the text of the conjecture and the name of the scholar that has made the proposal. A parser identifies these chunks of information and an aligment algorithm is applied to find the exact position in the verse where the conjectures is intended to be collocated.

## 3 Putting all Together

The search engine that will be developed, extending the model of [13], should allow the research of variants and conjectures in their contexts, showing the actual page of the manuscript where the variant is attested or the image of the printed edition where the conjecture was formulated.

## 4 Conclusions

Digital Aeschylus is an ongoing project focused on the textual tradition of a single ancient author. The first stage concerns the acquisition, digitization and linkage of the materials.

The second stage will concern the application of corpus analysis to the acquired documents. First experiments on the current corpus are promising, as illustrated in [4].

## References


1. Abbyy FineReader Homepage, http://www.abbyy.com
2. Anagnostis Homepage, http://www.ideatech-online.com
3. F. Boschetti: Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses onto Reference Text. Proceedings of the Corpus Linguistic Conference (27-30 July 2007)
   http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf
4. F. Boschetti: Gli Spazi Semantici del Greco Antico. (to appear in Quaderni Urbinati 2008)
5. F. Boschetti, M. Romanello, A. Babeu, D. Bamman, G. Crane: Improving OCR Accuracy for Classical Critical Editions. (to appear in ECDL 2009)
6. L. Burnard, S. Bauman: TEI P5 – Guidelines for Electronic Text Encoding and Interchange. Oxford (2008)
   http://www.tei-c.org/Guidelines/P5
7. G. Crane, D. Bamman, L. Cerrato, A. Jones, D. Mimno, A. Packel, D. Sculley, G. Weaver: Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. 10th European Conference on Research and Advanced Technology for Digital Libraries, volume 4172 of Lecture Notes in Computer Science, 353-366, Springer (2006)
8. G. Crane, B. Seales, M. Terras: Cyberinfrastructure for Classical Philology. Digital Humanities Quarterly, 3, 1, 1–27 (2009)
9. Internet Archive Homepage, http://www.archive.org
10. JStor Homepage, http://www.jstor.org
11. W.B. Lund, E.K. Ringger: Improving Optical Character Recognition through Efficient Multiple System Alignment. (to appear in JCDL 2009)
12. Multitext Homer Homepage,
    http://chs.harvard.edu/chs/homer_multitext
13. Musisque Deoque Homepage, http://www.mqdq.it
14. OCRopus Homepage, http://code.google.com/p/ocropus
15. Perseus Project Homepage,
    http://www.perseus.tufts.edu/hopper/opensource
16. S. Reddy, G. Crane: A Document Recognition System for Early Modern Latin. Chicago Colloquium on Digital Humanities and Computer Science: What Do You Do With A Million Books, Chicago, IL (2006).
17. M. Spencer, C. Howe: Collating texts using progressive multiple alignment. Computer and the Humanities, 37, 1, 97–109 (2003)
18. G. Stewart, G. Crane, A. Babeu: A New Generation of Textual Corpora. JCDL 2007, 356–365 (2007)
19. M. Taufer: Stato del New Repertory of Conjectures on Aeschylus e della Collezione di Manoscritti Eschilei. (to appear in Quaderni Urbinati 2009)
20. A. Wartelle: Bibliographie Historique et Critique d'Eschyle et de la Tragédie Grecque, 1518-1974. Paris (1978)

# Topic Classification Using Limited Bibliographic Metadata

Kerstin Denecke [1], Thomas Risse [2], Thomas Bähr [3]

[1,2]L3S Research Center, University of Hannover, Germany
[3]Technische Informationsbibliothek Hannover, Germany
[1]denecke@L3S.de

**Abstract.** In this paper, we introduce a method for categorizing digital items according to their topic, only relying on the document's metadata, such as author name and title information. The proposed approach is based on a set of lexical resources (e.g., journal titles, conference names), on terminology extraction and traditional machine-learning technologies. Evaluation results on a real world data set show that the approach achieves promising results.

## 1 Introduction

Specialized public libraries, such as the British Library or the German National Library of Science and Technology (TIB), and commercial information providers offer barrier-free access to resources and an optimized, user-oriented search interface. To narrow the information space for searching and browsing, knowledge about a document's topic is required that can be provided by controlled classifications schemes and index terms. Due to the monthly increasing amount of catalogue entries, manual classification of all data items is impossible; automatic technologies are required. The development of such methods is one of the objectives of the LinSearch project (http://www.linsearch.de) which is partly funded by the German Federal Ministry of Economics and Technology (BMWI). In this paper we present some of the results of this project.

The considered classification problem is to assign one class out of 14 possible classes to a single catalogue entry of the TIB data collection. The objective of this classification is to support the user and also the retrieval process in restricting search results to those belonging to the domain of interest. A more fine-grained classification is not useful in this case. Our work focuses on classes that are especially relevant for the TIB whose data collection consists of documents on technology and engineering. Possible classes are 'computer science', 'mathematics', 'physics', 'architecture', 'chemistry', 'engineering', 'civil-', 'chemical-', 'electrical-', 'power-', 'production-', 'mechanical-', 'environment-', and 'process engineering'. The TIB data collection consists currently of around 15 million entries and every month between 10,000 and 30,000 new items are integrated. Each catalogue entry represents a journal paper, conference paper, a book or a research report and comprises a set of metadata. Depending on the amount of metadata available for a data item, we distinguish four different levels of data quality. Data of level I only offer document and publication information. If in addition the journal or conference information is available, the data item belongs to quality level II. Data items

of level III offer an abstract, and those of level IV provide classification information. In case of TIB this is the so-called "Basis Klassifikation (BK)" (base classification system, [1]). This hierarchical decimal classification system was originally developed in the Netherlands and is mainly used by libraries in the Netherlands and Germany.

## 2 Related Work

The most commonly used text representation for topic classification is the vector space representation, where each distinct word in a document collection acts as a feature [2]. Other classification features include syntactic [3], semantic and stylistic features such as character or word sequence frequencies [4] or n-gram frequencies [5]. Statistical features on words (term frequency and the like) as mainly used by existing approaches are insufficient in our context, since within document titles term frequencies may not differ significantly and the number of topic-related terms is reduced. Hulth and Megyesi use extracted keywords as classification features [6]. In our approach, also keywords are exploited but they are extracted from title information only and integrate additional bibliographic data to the feature set. So far, there is no study available that considers topic text classification for these specific conditions. Montej-Rez et al. exploit metadata in combination with extracted keywords and a multi-label classifier TECAT for text classification [7]. The keywords are extracted from the document itself, which is in our task unavailable.

## 3 Classification Approach

Our approach to address the previously described problem combines rule-based classification with machine learning techniques. First, the metadata of a given catalogue entry is checked for an available classification that can be mapped to one of the 14 desired classes (data of quality level IV). For this purpose, mapping rules have been established manually to map from assigned codes of the base classification system [1]. Data items of quality level II and III are processed in the second step where class-specific information on journal titles and conference names are exploited to assign a class label. For this purpose, lists with conference and journal titles have been collected from existing class-specific repositories (e.g., from DBLP) as well as from already classified data items for each class under consideration. In case the first two steps fail due to missing information as well as for data items of quality level I, core features are identified in the metadata and used by a machine-learning classifier.

The final feature set for the classification consists of the following attributes: (1) the first five author names, (2) the publisher information, (3) the name of the corporate creator, and (4) a class-specific score for each category. Attributes 1 to 3 are directly derived from the metadata information. To calculate the scores (attribute 4) the document title and - if available - its abstract is exploited. A class-specific score corresponds to the number of matches of keyphrases extracted from a document and a class-specific term list. For each category under consideration such a class-specific term list of domain-relevant terms and phrases has been established semi-automatically from existing resources and from already classified material. Keywords and -phrases, i.e. word groups

with a maximum of 5 words that are neither stop words nor start or end with a stop word are extracted from title and abstract information. Each extracted keyphrase is looked up in the class-specific term lists; matches are counted per class resulting in a class-specific score per category.

The resulting feature set is exploited for document classification by a machine-learning classifier. Different algorithms that are implemented in the WEKA library [8] have been tested. The LogitBoost classifier based on logistic regression performed best and is therefore used in our experiments.

## 4 Evaluation

In this section, we describe the results when applying the introduced algorithm to the TIB dataset. In a first experiment, manually classified documents derived from different publishers and of quality level I or II (i.e., abstracts and BK-codes are unavailable) have been exploited. In a 10-fold-cross validation with 1500 documents per category the method achieved an accuracy of 86.7%. We also tested the algorithm with different feature sets and conclude that class-specific scores are well suited as features while exploiting additional metadata such as author names or publisher is not helpful in the given context. A possible explanation for this is that in the given data collection, documents of the same author are very rare.

In an additional evaluation, a second data set with unclassified data entries was manually evaluated. The classifier was trained on 1500 documents per category. Five employees of the TIB manually evaluated documents of their special field. Specialists for a the categories 'mathematics', 'mechanical engineering', 'chemistry', 'physics' and 'engineering' were involved. The other categories will be considered in future evaluations. From the 1180 data items, 82.7% were correctly classified by our approach. Documents of the domains 'chemistry' and 'physics' were almost completely correct classified (99% accuracy). Accuracy results of 70% and 87% were achieved for documents of the domains 'engineering' and 'mechanical engineering'. The worst results were achieved for the categories 'mathematics' (58% accuracy). For a more comprehensive description of the evaluation we refer to the full paper on our approach. Evaluation results of Level III and IV data will be represented in a different paper.

## 5 Conclusion

In this work, a text classification approach is introduced that relies only on bibliographic metadata. We show that despite this reduced semantic information good classification results are achieved. The best results are obtained when relying only on the title and abstract information. The lexical resources used by the approach can be easily extended and allow in this way an easy modification to similar classification problems. Furthermore, these term lists can be used to support the indexing process. In future work, we will test whether the assigned categories can be exploited to improve user satisfaction in document retrieval. For this purpose, the method will be integrated into the processes of the TIB to improve search facilities and support the restriction of search results.

## References


1. Common Library Network GBV: Basisklassifikation (2008) http://www.gbv.de/vgm/info/mitglieder/02Verbund/01Erschliessung/02Richtlinien/05Basisklassifikation/index.
2. Sebastiani, F.: Machine learning in automated text categorization. In: ACM Computing Surveys, 34(1), Kluwer Academic Publishers (2002) 1–47
3. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: Proceedings of the 26th European Conference on Information Retrieval Research. (2004) 181–96
4. N. Cancedda, E.G., Goutte, C., Renders, J.: Word sequence kernels. In: Journal of Machine Learning Research. (2003) 1059–82
5. Peng, D., Schuurmans, F., Wang, S.: Language and task independent text categorization with simple language models. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. (2003)
6. Hulth, A., Megyesi, B.B.: A study on automatically extracted keywords in text categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. (2006) 537–44
7. Montejo-Raez, A., Urena-Lopez, L., Steinberger, R.: Text categorization using bibliographic records: Beyond document content. In: Procesamiento del Lenguaje Natural, 35. (2005) 119–262
8. Witten, I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)

# Fostering User Experience in order to Improve the Quality of a Digital Library

Laurent Eilrich[1], Frederic Andres[1,2], Ghislain Sillaume[1], and Marianne Backes[1]

1 CVCE, Chateau de Sanem,
L-4992 Sanem, Luxembourg
{laurent.eilrich, frederic.andres,ghislain.sillaume,marianne.backes}@cvce.lu
2 NII, Tokyo, 101-8430, Japan
andres@nii.ac.jp

**Abstract.** We present our approach to the redesign of a Digital Library (DL) related to the history of European integration (called European NAvigator, ENA). For the next version of this DL (ENA 2010), special attention has been paid to improving the user experience. We consulted with DL user communities before starting the redesign and obtained a list of users' expectations and needs regarding the DL. Although some expectations and needs could be met by putting certain functionalities on the screen; others were more difficult to meet because they went beyond the expected role of a DL, which is to facilitate access to objects of information and offer a collection of services to users. Our study demonstrated the utility of conducting a users' study in the preliminary phases of a DL project. Such a user-centered design approach enables users' perspectives to be incorporated into the DL.

**Keywords:** Digital Library, ENA 2010, User Experience, Usability, User-Centered Design Approach.

## How to foster User Experience in ENA ?

The digital revolution has been the impetus for a lot of work on digital libraries. Over the last fifteen years, there have been theoretical studies on digital libraries [5], and digital libraries have grown up thanks to technological innovations of IT specialists. In particular, one library, called European NAvigator[1], has evolved from a CD-Rom into a multimedia Internet application through a client/server application broadcasted by satellite. The evolution of this library reflects a succession of technical opportunities [6]. Leiner [12] defines a DL to be a collection of services and information objects that support users in dealing with information objects by organizing and presenting them directly or indirectly via electronic/digital means. ENA is a DL that provides high-quality research and educational materials about the history of European governmental and social integration. It is a multilingual, multi-source, and multimedia knowledge base that contains more than 15,000 documents about the historical and institutional development of a united Europe from 1945 to the present day. Students, teachers, researchers, and anyone interested in the history of European integration can find original materials including photos, audio, video clips, press articles, and cartoons, together with explanatory synopses, tables, interactive maps, and diagrams. Each ENA material has been selected and scrutinized by a

[1] ENA: www.ena.lu

multidisciplinary team of specialists in European integration. Specialists select relevant sources from a large variety of documentary sources (e.g. publishing houses, periodicals, historical archives, organizations, public institutions, and other bodies) and apply strict selection criteria to ensure the quality of each ENA information object. Its documents form a part of the European heritage. Furthermore, it contains many supplementary materials, e.g., rare, unpublished, and difficult-to-access materials, relevant to the theme of European integration. These materials illustrate how and why Europeans have banded together within this union (or have chosen to reject it). ENA takes into account current research issues, especially historiographic methodologies. For example, the choice of materials included in the European Organizations section is based on precise criteria, e.g. legal instruments and learned articles.

The goal of our project has been to take this reliable knowledge base on the history of European integration and create a new version called ENA 2010 that will provide its user communities with convenient and useful web 2.0 technologies.

Fuhr et al. [8] believe that the starting points of designing a DL are determining its intended usage and corresponding user needs. Mahlke [14] shows that research into user's needs, etc., is valuable for defining the product. Like Fuhr et al., he believes that identifying the users' needs is a starting point to developing an innovative system. Beringer and Holtzblatt [2] emphasize the necessity to clarify the objectives of user research. Xie [17] pointed out that incorporating users' perspectives into the development of digital libraries requires the users' perceived importance of the DL evaluation criteria, knowledge of how they use digital libraries, and their evaluations of digital libraries, as well as their preferences, experience, and knowledge structures. Xie also stresses the importance of different evaluation criteria for different categories of users: i.e., users, researchers and professionals. Lettl et al. [13] noted that certain categories of users contribute more to the development of radical innovations.

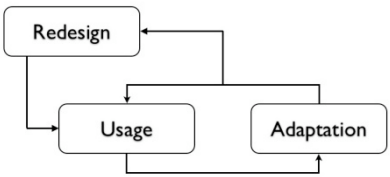



Figure 1: Usage-adaptation-redesign cycle

We should learn the users' needs and how they use a DL. Brangier [3] proposes the usage-adaptation-re-engineering cycle (see Figure 1), which highlights how human adaptations (of the users) are a source of innovation to design new uses. This idea of design is related to the question of use, which requires the existence of a DL or a prototype. The concept of innovation from current use is also discussed by Dix [7], who suggests the appropriation concept as a source for innovation in interactive system design. An appropriation exists when an object is used in another way that the way the designer has imagined. The models suggested by Brangier [3] and Dix [7] lead us to look at actual use cases in order to define users' needs precisely.

In the following, we will focus on the utility of a users' study in the preliminary phases of a DL project. To develop the next version of the DL, we decided to study

user communities as a first step in a user-centered design approach study. After that, we focused on DL usability. Indeed, since usability is an essential asset for any product [14], we knew that the success of the next DL depends not only on its technical qualities, its design, and its robustness, but also on its usability. Jeng [9][10] supports this idea by proposing an evaluation model of usability based the original criteria of the standard ISO 9241: Effectiveness, Efficiency, Satisfaction, and Learnability. This evaluation model of usability requires a use and thus the existence of a DL or a prototype. The model for the evaluation of the usability of the systems of digital libraries has been revisited by Tsakonas and Papatheodorou [16]. They showed that usability and usefulness are the two major criteria to evaluate DLs. These two criteria are essential to create an overall positive user experience. Hence, we decided to focus on usability during the DL development process by evaluating the usability of each prototype of ENA 2010. The ENA user communities' study was conducted by Brangier, Dinet, and Eilrich [4] and with the support of the ETIC[2] laboratory (User Experience Laboratory).

The first phase consisted of determining the conditions of use and the technological and informational profiles of the potential users. We identified communities of users likely to use our DL. We defined fourteen communities of practice involved in the European Integration Process. These included communities of researchers/historians, lawyers, professionals in documentation, journalists, teachers, international teachers, students, PhD students, software ergonomists, politicians, computer engineers, experts in intercultural studies, experts in new digital leisure, and members of historical associations.

The second phase consisted in an evaluation of DL usability by using the classical ergonomic criteria of Bastien and Scapin [1]. The third phase was an analysis of user requirements and an investigation of DL uses. We met with members of each community during 14 video-recorded focus groups. Each idea expressed by the groups concerning a need or a use could be listed. Altogether, 53 new ideations were identified. After classifying them, we realized that some could be supported by putting functionalities on the screen, whereas others were too complex because they went beyond the original role of a DL, which is to facilitate the access to objects of information and offer a collection of services to the users. Indeed, the idea of *having a direct phone contact* was expressed by a group of journalists. We noticed that a specific user need is hard to fulfill because it involves creating a new job in the DL, in order to take care of users. Xie [17] made similar observations, and some users suggested as DL evaluation criterion the unique services offered by Kim's model [11]. It gave precise details related to this topic in relationship with ours: DL qualities affect user satisfaction, DL use, and organizational impact.

The DL user communities' study gave us a clear vision about our users' needs. It was a good starting point to create a user-centered DL. Furthermore, we suppose that each functionality affects usefulness and usability and that it is necessary to appreciate these elements in an objective way by using the usability model proposed by Jeng [9]. The various services will have to conform to users' expectations. A clear model of a user-centered digital library design will have to support its own evolution in order to provide a positive user experience.

---

[2] ETIC : Equipe Transdisciplinaire d'Interaction et de Cognition

**Acknowledgments**

We would like to thank the CVCE for its support during the ENA 2010 project.

## References

1. Bastien, J. M. C., & Scapin, D. L., 1993, Ergonomic criteria for the evaluation of human-computer interfaces (Report No. 156). Rocquencourt, France: Institut National de Recherche en Informatique et en Automatique
2. Beringer, J. & Holtzblatt, K. (2006), *Designing Composite Applications*, Palo Alto, CA: SAP Press
3. Brangier, E., (2004), La boucle usage adaptation - reconception : l'usage comme intégration des points de vue de l'utilisation et de la conception, In P. Rey, E. Ollagnier, V. Gonik and D. Ramaciotti, Ergonmie et normalization, Toulouse : Octares, Collection le travail en debats, pp 535-544.
4. Brangier, E., Dinet, J.& Eilrich, L. (2009). The Seven Basic Functions of a Digital Library. Conference HCII, 2009, pp 345-354
5. Candela et al. (2008), The DELOS Digital Library Reference Model – Foundations for Digital Libraries". Technical Report. DELOS.
6. Consalvi M., Eilrich L., Sillaume G. (2008), Rôles et perspectives des bibliothèques numériques dans l'économie de la connaissance : une étude de cas, Colloque En route vers Lisbonne. Luxembourg . To be published in the journal "Perspectives de politique économique".
7. Dix, A. (2007). Designing for Appropriation. In *Proceedings of the 21st BCS HCI group conference* (volume 2).
8. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., and Sølvberg, I. (2007), Evaluation of digital libraries. *Int. J. Digit. Libr.* 8, 1 (Oct. 2007), 21-38.
9. Jeng, J. (2004). "Usability of Digital Libraries: An Evaluation Model," jcdl, pp.407-407, Fourth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'04), 2004
10. Jeng, J. (2005). What is usability in the context of the digital library and how can it be measured?" *Information Technology and Libraries 24*(2), June 2005, p. 47-56.
11. Kim K. (2002). A Model-based Approach to Usability Evaluation for Digital Libraries. JCDL'02 Workshop on Usability of Digital Libraries Usability of Digital Libraries.
12. Leiner B.M. (1998). The Scope of the Digital Library. DLib Working Group on Digital Library Metrics http://www.dlib.org/metrics/public/papers/dig-lib-scope.html
13. Lettl, C., Herstatt, C., Gemuenden, H.G. (2006): Learning from users for radical innovation. *International Journal of Technology Management*, Vol. 33 (1), 25-45.
14. Mahlke, S. (2008). Learning from Users to Innovate: User Research and Innovation. Proceedings upaEurope2008
15. Marcus, A. (2002). Return on Investment for Usable User-Interface Design: Examples and Statistics, Version 26 February 2002
16. Tsakonas, G. & Papatheodorou, C. (2006), Analyzing and evaluating usefulness and usability in electronic information services. *Journal of Information Science 32*(5), pp. 400-419.
17. Xie, H. I. (2008), Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. *Inf. Process. Manage.* 44, 3, 1346-1373.

# Accessing Media Art via the GAMA Portal

Björn Gottfried[1], Andree Lüdtke[1], Gaby Wijers, Anna-Karin Larsson, Ida Hirsenfelder, Eva Kozma, and Otthein Herzog[1]

## 1 Introduction

This paper presents work on how *media art* has found its way into digital libraries in order to be accessible for a broad community. Digital libraries provide comfortable means for collecting, managing, preserving, and distributing digital content. In the case of media art this is of interest for curators, artists, academics, researchers, mediators, and for the interested public.

A number of challenges arise when making available media art in digital libraries. These challenges include in particular the question how to search within content of media art that consists of a large number of art videos – the analysis of the latter being reviewed in [5]. In [6] it is shown how metadata about video content is employed. On the one hand, metadata is directly imported from the databases of content providers who annotate their videos and images with information; it is then the challenge to integrate data from different sources with heterogeneous data models and to ensure interoperability. On the other hand, content-based metadata is extracted from the raw media content, such as descriptions of audiovisual characteristics of media content; this is typically not available from the content providers and extracted by a content-based indexing service automatically. Both kinds of metadata can then be used in order to search in media art content.

A web portal has been developed for accessing several European media art collections that deploys the aforementioned search capabilities [4]. While [6] provides the technical background in order to deal with the underlying metadata search-engine, here we focus on the ideas behind this portal, how it presents itself to the user, outline which content is provided and what the users of this portal can expect from its use, i.e. its purpose and functionality.

### 1.1 Structure

The final paper will be structured in the following way. After a general introduction, a motivation part presents the philosophy behind this media art portal. The next section describes the portal itself and a list of requirements from the point of view of the users. Then, those requirements are picked up to show how the web portal supports the users according to their requirements. An outlook in the end points out a couple of future challenges. In the following, we summarise these issues.


**Fig. 1.** A screenshot of the GAMA web portal at http://gamaweb.hku.nl/.

## 2 Libraries of Media Art

The number of libraries who are actively involved in building digital repositories is growing. A large body of content deals with books, journals, papers, and other works. Another more specific but also growing segment of libraries deals with media art. In fact, media art is currently becoming one of the most popular contemporary art genres. Bringing together culture and technology it is natural to employ the word wide web in order to present and access media art. Related work can be found in [7, 8, 1, 2].

### 2.1 The GAMA Portal

Fig. 1 shows the GAMA portal. It is the aim of the portal to establish a central platform to enable the access to media art archives, that is their digitised contents. The archives which are involved so far comprise a majority of the most important digital content holders for media art in Europe. It is the idea to ensure a significant increase in use, re-use and cross-border visibility of the digital content when aggregating such a large amount of media art content, and also, to be

able to access this spatially distributed content through one common interface. The integration of GAMA into the broader community of the cultural heritage is planned for the future via the Europeana portal [3].

## 3 Objectives and expected Outcomes

The addressed objectives of the GAMA portal are as follows:

- Enable enhanced access to European media art archives.
- Significantly increase awareness and mediation of media art.
- Be Europe's key online portal to its media art archives.
- Facilitate to users the discovery, use and re-use of European digital cultural and artistic contents in Europe.
- Combine and adapt existing standard and state-of-the art solutions to meet the needs for interoperability between the individual archives and their heterogeneity.

The expected outcomes of the GAMA system include:

- A sound framework and quality procedures for media annotation in order to be able to expand the GAMA gateway content and activities also to other content types.
- Internet and systems compatibility of content by using existing standards and respective logistics for server operations.
- Integration of state-of-the-art automatic metadata indexing and video segmentation tools in order to provide fast access and content browsing capabilities.
- Online availability of the GAMA gateway with an operable and advanced user interface with web accessibility standards.
- Advanced search facilities (like image query by example and visual similarity search) combined with keywords to ease the finding of media art items.

## Acknowledgments

The GAMA project is funded under the eContentplus programme of the European Commission; Grant No. ECP-2006-DILI – 510029. The authors would like to thank the whole GAMA consortium.

## References

1. Centre Pompidou. New media encyclopedia. Website, 2009. Available online at http://www.newmedia-art.org visited on 25 May 2009.

2. Daniel Langlois Fondation. Explore artworks, themes, artists, publications. Website, 2009. Available online at http://www.fondation-langlois.org/html/f/ visited on 25 May 2009.
3. Europeana. Europeana: connecting cultural heritage. Website, 2009. Available online at http://dev.europeana.eu/home.php visited on 16 June 2009.
4. GAMA Consortium. Gateway to archives of media art. Website, 2009. Available online at http://gamaweb.hku.nl/ visited on 5 June 2009.
5. S. Lefevre, J. Hollera, and N. Vincent. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, 9:73–98, 2003.
6. A. Luedtke, B. Gottfried, O. Herzog, G. Ioannidis, M. Leszczuk, and V. Simko. Accessing Libraries of Media Art through Metadata. In R. Chbeir, K. Coninx, F. Ferri, and P. Grifoni, editors, *3rd International Workshop on Management and Interaction of Multimedia Information Content*. To appear at IEEE Press, 2009.
7. Ubu Web. UBU Portal. Website, 2009. Available online at http://ubu.clc.wvu.edu/ visited on 25 May 2009.
8. VideoArtWorld, S. L. Art in movement. Website, 2009. Available online at http://www.videoartworld.com/beta/home.php visited on 25 May 2009.

# The LivingKnowledge Project: Exploring the Spectrum of Opinions over Time

Richard Johansson and Alessandro Moschitti

DISI, University of Trento
Trento, Italy
{johansson, moschitti}@disi.unitn.it

**Abstract.** The last two decades of research in Information Retrieval have shown that *bag-of-words* models are sufficient for the design of document search and categorization systems. This has made useless, at least for document retrieval purposes, the development of semantic models more advanced than the simple *bag-of-words*. In contrast, recent research in sentiment classification has shown that, when the required semantic information is not limited to query-document relatedness, e.g. opinion mining, advanced semantic processing is crucial.
In this perspective, the LivingKnowledge (LK) project, funded by the seventh EU framework program, aims at studying and developing a technology based on semantic processing, which can be exploited to solve complex semantic tasks such as opinion extraction, opinion analysis in terms of diversity and their evolution over time.
The role of LK's work is twofold: (a) it is fundamental for the design of innovative future digital libraries since different opinions can be other dimensions for searching or categorization of the digital content and (b) the evolution of opinions also refers to the study of the evolution of knowledge and categorization schemes during time. This document gives an overview of the two main objectives of the project.

## 1 Introduction

LivingKnowledge (LK) is a project on future emerging technology, funded by the seventh EU framework program. Among other its research subjects, e.g. design of automatic tools for social science analysis, LK aims at studying and developing semantic processing models for opinion extraction, opinion analysis and knowledge evolution over time.

The first two aspects are rather interesting for digital library research since the automatic extracted metadata like for example: `opinionated` or `pos./neg. opinion`, allows for searching or categorizing documents according to standard topics as well as this new semantic dimension. For example, we can categorize films based on genres: *adventure*, *dramatic*, *horror* and so on along with the polarity of opinions[1]. The latter can be also characterized with a finite interval of values, e.g. from 1 to 5.

[1] Another interesting and related task is the product review mining [1]

Work on sentiment classification [2] has shown that, in contrast to standard text categorization [3, 4], syntactic/semantic processing is required to boost the performance of the *bag-of-words* models. In this perspective, LK will explore the most advanced technology for encoding syntax and semantics, i.e. support vector machines [5] based on structured kernels, e.g. [6], for encoding syntactic parse tree information along with predicate–argument structures [7–9] (semantic structures) in the automatic opinion analyzers.

Regarding knowledge evolution, for which opinion dynamics is just an instance of knowledge, the project is studying: (a) ways to adapt categorization systems to the evolution of document content over time such data they maintain a satisfactory accuracy; and (b) approaches to the scheme evolution so that categories are automatically created, deleted or merged.

In the remainder of this paper for sake of space we will only illustrate the opinion mining aspects along with our preliminary results on simple models and the planned advanced approaches.

## 2 Automatic Retrieval of Opinionated Pieces of Text

Automatic retrieval of opinionated pieces of text may be carried out on a number of different levels. On the most coarse level, *documents* are categorized as opinionated or factual; for instance, this may be used to distinguish editorials from news [10].

At the other end of the spectrum, methods have been proposed to carry out fine-grained subjectivity analysis on the level of linguistic expressions [2].

In the ongoing project, we currently focus mainly on the automatic classification of individual *sentences* as opinionated or not. This will later pave the way for a more fine-grained analysis that can support a detailed exploration over time of the opinions held by various groups of people.

### 2.1 Preliminary Experiments in Sentence-level Opinion Classification

As a first step, we formalized the problem of detecting opinionated sentences as a binary text categorization problem. The problem could then be approached using classical statistical text categorization techniques. We thus represented a sentence as a vector in a high-dimensional space using a bag-of-words representation and trained a binary statistical classifier to distinguish the two types of sentences (subjective or objective). The classifiers were linear support vector machines, which have previously been shown to be effective in text categorization problems [5]. To train and evaluate the classifier, we used the MPQA corpus, a collection of 692 documents in English (containing 15,768 sentences) manually annotated with information about expressions of subjectivity [2].

In addition to the classifier based on a pure bag-of-words representation, we implemented a classifier that also used a subjectivity lexicon to determine the presence of strongly or weakly subjective words in the sentence (such as

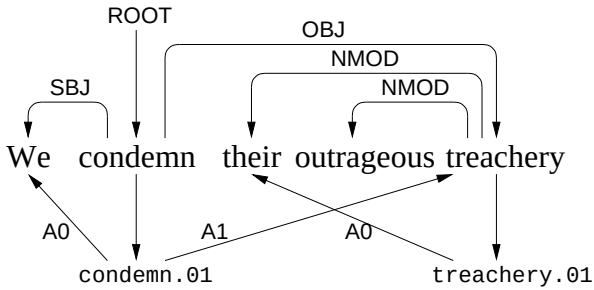



**Fig. 1.** Example of a syntactic–semantic dependency graph.

*wonderful*, *condemn*). We evaluated the classifiers for subjective sentences on a subset of 2,185 sentences of the MPQA corpus by obtaining a precision, a recall, and an F1-measure of 0.79, 0.76 and 0.78, respectively. When we used the lexicon the figures improved to 0.82, 0.79 and 0.81, respectively.

### 2.2 Future Work on Linguistic Structure for Opinion Extraction

It is still an open question which linguistic information is useful for the automatic retrieval of opinionated sentences. Previous methods for finding opinions have relied either on on simple cues or keyword spotting [11] or on bag-of-words methods [12], as described in the previous section.

We hypothesize that deeper linguistic structures may be useful for opinion retrieval, and we will explore various linguistic representations as a part of this research. As a start, we will see whether it is possible to use automatic syntactic and role-semantic analysis of sentences to improve the classifiers similarly to the approach followed for Question/Answer classification [6].

As an example, Figure 1 shows the analysis of the sentence *We condemn their outrageous treachery*. The sentence was automatically analyzed by the LTH parser [13]. In the figure, the syntactic representation is shown above the sentence and the semantic representation below. For instance, the syntactic graph shows that *We* is a syntactic subject of *condemn*, and the semantic graph shows that *their* has the A0 (BETRAYER) semantic role in the event represented by the word *treachery*. Such structure can easily be encoded in SVMs by means of structured kernels, [6], which represent it in terms of all its substructures (i.e. each feature is a portion of the graph). In addition to the syntactic and role-semantic graphs, we plan to explore other types of linguistic representation such as discourse graphs [14].

To conclude, we believe that the LK research will help to advance the research on digital libraries on three different lines: (i) the opinion classification will provide automatic metadata, which can refine the categorization schema and the access methods, (ii) advanced syntactic/semantic representation for opinions will provide theory and methods for the representation of other digital content,

e.g. definitions, explanations, and (iii) the study of the evolving categorization schema is directly related to the evolution and management of future digital libraries.

## References


1. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Edmonton, Canada (2002) 341–349
2. Wilson, T.A.: Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. PhD thesis, University of Pittsburgh, Pittsburgh, United States (2008)
3. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In McDonald, S., Tait, J., eds.: Advances in Information Retrieval – ECIR, Sunderland, UK. (2004)
4. Basili, R., Moschitti, A., Pazienza, M.T.: A text classifier based on linguistic processing. In: Proceedings of IJCAI 99, Machine Learning for Information Filtering. (1999)
5. Joachims, T.: Learning to Classify Text using Support Vector Machines. PhD thesis, University of Dortmund, Dortmund, Germany (2002)
6. Moschitti, A.: Kernel methods, syntax and semantics for relational text categorization. In: Proceeding of CIKM '08, NY, USA (2008)
7. Giuglea, A.M., Moschitti, A.: Knowledge Discovery using Framenet, Verbnet and Propbank. In Meyers, A., ed.: Workshop on Ontology and Knowledge Discovering at ECML 2004, Pisa, Italy (2004)
8. Moschitti, A., Pighin, D., Basili, R.: Tree kernels for semantic role labeling. Computational Linguistics **34**(2) (2008) 193–224
9. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J.: The CoNLL–2008 shared task on joint parsing of syntactic and semantic dependencies. In: CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, Manchester, United Kingdom (2008) 159–177
10. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), Sapporo, Japan (2003) 129–136
11. Wiebe, J., Bruce, R., O'Hara, T.: Development and use of a gold standard data set for subjectivity classifications. In: Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics. (1999)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP. (2002)
13. Johansson, R., Nugues, P.: Dependency-based syntactic–semantic analysis with PropBank and NomBank. In: CoNLL 2008: Proceedings of the Twelfth Conference on Natural Language Learning, Manchester, United Kingdom (2008) 183–187
14. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B.: The Penn Discourse Treebank 2.0. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). (2008)

# DISMARC and BHL-Europe: multilingual access to two aggregation platforms for Europeana

Walter Koch[1], Henning Scholz[2]
[1] AIT Angewandte Informationstechnik Forschungsgesellschaft mbH, Klosterwiesgasse 32, 8010 Graz, Austria
[2] Museum für Naturkunde – Leibniz Institute for Research on Evolution and Biodiversity at the Humboldt University Berlin, Invalidenstrasse 43, 10115 Berlin, Germany

**Abstract.** Digital audio content and digitized biodiversity literature are aggregated in two platforms and delivered to Europeana, the European Digital Library. The audio platform which is already fully operational was developed in course of the DISMARC project and provides the baseline for multilingual data mapping and access modules of Biodiversity Heritage Library (BHL) for Europe, an eContentPlus project which complements the successful BHL operations started in the United States. Multilingual vocabularies which are exposed as web services are used for semantic enrichment of data during the input process, for the query expansion and when presenting search results.

**Keywords:** DISMARC, audio, BHL-EUROPE, biodiversity, SKOS, WebServices, semantics, Europeana

## 1 Introduction

Within the European Digital Library (Europeana)[1] concept the access to domain specific or national/regional aggregation platforms is of paramount importance. Several projects funded by the European Commission through the eContentPlus Programme are targeting to set up aggregation systems which provide metadata to Europeana and links to digital resources. An "audio-pillar" for the Pan-European Library has been developed during the implementation of the DISMARC (DIScovering Music ARChives)-Project which started in 2006 end finished successfully in August 2008. After half a year fully operation the DISMARC[2] meta store (aggregation platform) contains nearly 2 million metadata records in different languages. Due to the fact that the project was lead by the "Multi-Kulti" department of RBB (Radio Berlin Brandenburg) it was possible to develop multilingual vocabularies and word lists in over 25 languages since for all these languages

[1] Europeana – a single access point to Europe's cultural heritage (15 June 2009), http://ec.europa.eu/information_society/activities/digital_libraries/europeana/index_en.htm
[2] DISCovering Music ARChives (15 June 2009), http://www.dismarc.eu

translators have been available. The vocabularies, implemented as SOAP[3]/WSDL[4] based WebServices, are used during the harmonization and input processes of archival records as well as at the user side when accessing the DISMARC meta store via the multilingual query and presentation interface. The DISMARC technical system will be further enhanced for the management of audio content related metadata within the EuropeanaConnect[5] project and forms the basis for a first pilot system supporting the Europeana Aggregation Platform for biodiversity literature. This work is carried out within the framework of the BHL-Europe project which can be considered as an "European Complement" to BHL[6], the Biodiversity Heritage Library, and started as a three year eContentPlus project in May 2009.

## 2 The DISMARC audio aggregation platform for Europeana

As stated in the Description of Work: "DISMARC uncovers large amounts of under-exposed European cultural, scientific and scholarly music audio. Content providers archives, broadcasters, museums, universities, research institutes, private collectors will be able to open up their collections to the wider world"[1], the project acts as an "audio aggregator" and guarantees operations for five years beyond the official project end which was in autumn 2008. The technical components of the DISMARC (DM) system consist of the "DM-meta store node" and the "DM-ontology node", both of them can be accessed via the DM-portal at www.DISMARC.eu. DM nodes offer functionalities for managers, domain experts, and end users either in the "back office" or "front office" mode of the portal; the different roles of DM actors within the different processes (input, aggregation, data mapping, vocabulary management, data access, etc) have been defined in the "DM-workflow" which has been elaborated using BPMN[7] the Business Process Modelling Notation.

### 2.1 The DISMARC nodes

The DM meta store node can be a constituent part of a "DM aggregation network" functioning as "DM-sub node" to another "DM-meta store node" and can be delivered in SaaS[8] - Software as a Service - mode as image of a Virtual Engine ("DISMARC-on-a-Stick"). This node includes sub components like: OAI[9] provider and harvester,

[3] SOAP Version 1.2 Part 1: Messaging Framework (Second Edition) (15 June 2009), http://www.w3.org/TR/soap12-part1/

[4] Web Services Description Language (WSDL) Version 2.0 (15 June 2009), http://www.w3.org /TR/wsdl20

[5] EuropeanaConnect (15 June 2009), http://www.europeanaconnect.eu/

[6] BHL-Wiki (15 June 2009) https://bhl.wikispaces.com/

[7] Business Process Modeling Notation (BPMN) (15 June 2009), http://www.omg.org/spec/ BPMN/1.2/PDF/

[8] Software as a service (15 June 2009), http://en.wikipedia.org/wiki/Software_as_a_service

[9] Open Archives Initiative (15 June 2009), http://www.openarchives.org/

search and browse subsystem (browsing supported by index lookup for all meta data elements or controlled vocabularies implemented as lists or trees (taxonomies), query expander, data mapping tools, administration and user management subsystem.

The DM ontology node offers import and export facilities based on SKOS[10] the Simple Knowledge Organisation System, Management tools (including translation services) for multilingual Controlled Vocabularies (CV) and SOAP/WSDL based WebServices which provide generic functionalities as described in ANSI Z39.19-2005[11] and Thesaurus specific functionalities (eg specific meta data elements which can be returned in a "query response" of a service); as registered user one can have access to the DM vocabulary WebServices (WS) via the DM-portal and integrate them into other applications. The vocabulary WS calculate "e-points" which can be used for charging service requests on a "pay per view" basis.

### 2.2 Multilinguality

Multilingual aspects are provided at different levels: the translators can expand a multilingual word list which contain preferred terms used in partner archives and the audio domain, multilingual vocabularies (eg IconClass[12]) can be imported, during the input and mapping processes of raw data provided by an Archive can be expanded by adding relevant terms taken from a DM controlled vocabulary, queries can be expanded by adding terms (via DM controlled vocabularies) in selected languages to search terms, result records show all equivalents of a term in translated languages provided the term is included in a DM CV, the portal language itself can be selected in 20+ languages.

## 3 BHL-Europe as aggregation platform for biodiversity literature

The Biodiversity Heritage Library (BHL) began as a consortium of 10 natural history, botanical, and research libraries in 2007, working together to digitize the published literature of biodiversity held in their respective collections and to make that literature available for open access and responsible use as a part of a global "biodiversity commons." Discussions are now underway with other nations moving BHL to a global initiative. BHL-Europe has recently started to make the biodiversity knowledge of Europe available to everybody who is interested by improving the interoperability of European biodiversity digital libraries. As a Best Practice Network, this will be done by the innovative application of proven technologies. Resources of other projects like DISMARC will be re-used to not reinvent the wheel. Eventually,

[10] SKOS Simple Knowledge Organization System (15 June 2009), http://www.w3.org/2004/02/skos/

[11] ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies (15 June 2009), http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a

[12] IconClass (15 June 2009), http://www.iconclass.nl/

BHL-Europe will provide a multilingual access point for digital content through EUROPEANA providing the first major corpus of science material to the European Digital Library. In addition, it will provide a robust and multilingual biodiversity community portal with sophisticated search tools and open, distributed architecture. This will be done in close collaboration with BHL to also support the internationalization of this initiative.

### 3.1 Technical aspects

The technical architecture of BHL-Europe is built around an OAIS [13]-compliant repository system. Within the Pre-Ingest processes mapping tools derived from the DISMARC project will be tested and integrated into the first BHL-Europe Pilot System which is foreseen to be available by end of 2009. It will also be checked if ETL[14] – (Extract-Transform-Load) technology as used in BI[15] (Business Intelligence) applications can be applied. The "BHL-Europe Access System" will take input from the existing BHL[16] System, specifications from BHL-Europe partners and the code base and experiences provided by DISMARC. Main activities for the first prototype contain: integration of a meta data scheme fulfilling the requirements of BHL, integration of a gateway for different Vocabulary WebServices (uBio[17], DISMARC, etc), data mapping for selected BHL-Europe partners and wrapping records for METS[18] based bulk load (ingest) into the BHL-Europe repository system, setup of an initial version of the BHL-Europe aggregation platform for Europeana. Further versions of the BHL-Europe systems will provide navigation in semantic networks which implementation is foreseen using XTM-TopicMap[19] standard eventually on top of a RDF triple engine (to be defined in course of the project implementation).

Since August 2009 a first prototype of the BHL-Europe test portal is available[20] and already provides access to 40.000 bibliographic items and connected digital resources from several BHL-Europe project partners. The following screen shots demonstrate the use of a multilingual thesaurus (eras) during query formulation and when presenting detailed metadata for a relevant item. This vocabulary (dmEras) was taken from the DISMARC project and is used to extend the metadata during the import process into the aggregation platform.

---

[13] ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model (15 June 2009), http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

[14] Extract, transform, load (15 June 2009), http://en.wikipedia.org/wiki/Extract,_transform,_ load

[15] Business intelligence (15 June 2009), http://en.wikipedia.org/wiki/Business_intelligence

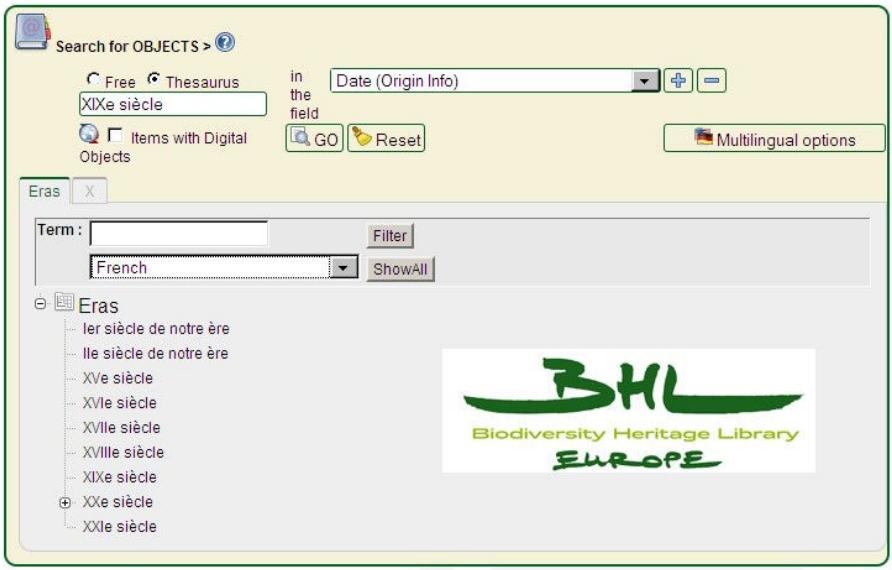
[16] BHL-System (15 June 2009), http://www.biodiversitylibrary.org/

[17] uBio (15 June 2009), http://www.ubio.org/

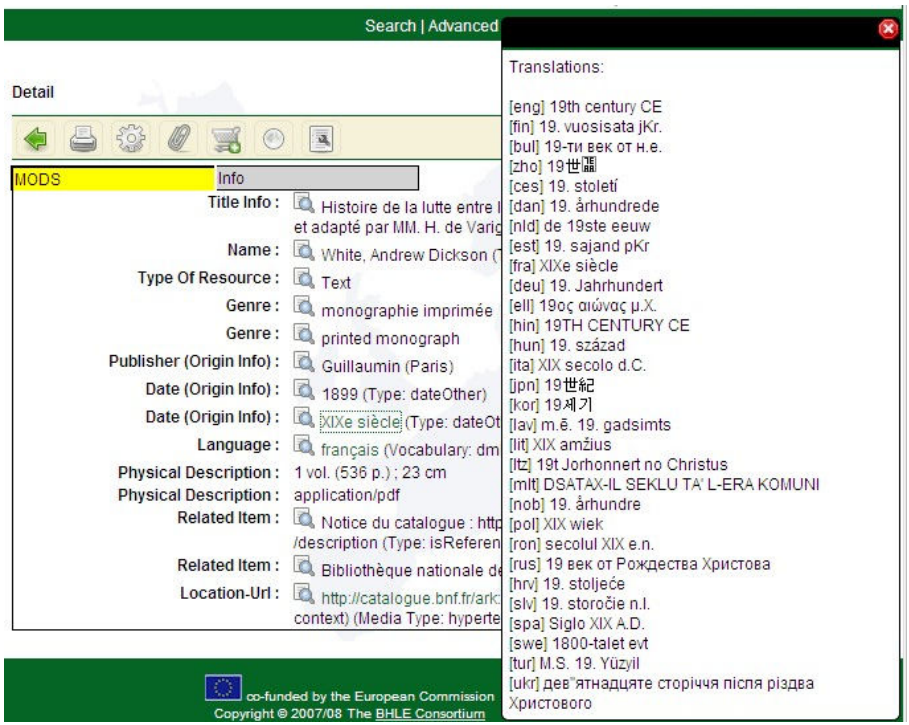
[18] METS (Metadata Encoding & Transmission Standard) (15 June 2009), http://www.loc.gov/standards/mets/

[19] XML Topic Maps (XTM) 1.0. (15 June 2009), http://www.topicmaps.org/xtm

[20] BHL-Europe; Biodiversity Heritage Library Test Portal (25 August 2009), http://bhl.ait.co.at




Screenshot 1: Thesaurus supported selection of search terms




Screenshot 2: Translation of the search term used into several languages

# Moving towards Adaptive Search

Stephen Dignum[1], Yunhyong Kim[2], Udo Kruschwitz[1], Dawei Song[2], Maria Fasli[1], and Anne De Roeck[3]

[1]University of Essex, Colchester, UK
{sandig, udo, mfasli}@essex.ac.uk
[2]Robert Gordon University, Aberdeen, UK
{ykim1, d.song}@rgu.ac.uk
[3]Open University, Milton Keynes, UK
a.deroeck@rgu.ac.uk

**Abstract.** Information retrieval has become very popular over the last decade with the advent of the Web. Nevertheless, searching on the Web is very different to searching on smaller, often more structured collections such as intranets and digital libraries. Such collections are the focus of the AutoAdapt project[1]. The project seeks to aid user search by providing well-structured domain knowledge to assist query modification and navigation. There are two challenges: acquiring the domain knowledge and adapting it automatically to the specific interest of the user community. The paper introduces an implemented prototype that serves as a starting point on the way to truly adaptive search.

## 1 Introduction

Document retrieval systems have been around for more than fifty years, and early systems exploited similar structures that we have in modern digital libraries, such as author name, book title, and keywords [9]. More recently we have witnessed a major shift towards search on the "Web". However, search techniques that work well on the Web do not necessarily work equally well on collections that have other characteristics, e.g. domain-specific or more structured collections, as found in intranets, digital libraries and on local Web sites. Retrieval from intranets, for example, behaves unlike Web search [14]. For instance, standard ranking functions (e.g., PageRank [3] and HITS [7]) that work well for Web collections are less effective on intranets. Furthermore, the terminology, structure, and services provided within such collections are selected to meet organisational requirements, and, consequently, a considerable amount of time is spent by users trying to learn the domain characteristics even before they are able to identify the adequate questions to be submitted to a search system. From an information systems perspective, thesauri and classification schemes should be developed and adapted to match information contained in such collections [1].

The approach that the AutoAdapt project takes is to maintain (or adapt) such structures automatically. We are looking at search as well as navigation within domain-specific document collections and our aim is to satisfy a user's information request effectively by learning from the entire user population and incorporating

[1] http://autoadaptproject.org

this learned knowledge in a constantly adapting domain model which assists a user in the search process. To support such adaptation we investigate how domain models are explored using clickthrough data that link up query modification steps and associate clicked documents with queries. This provides a context-rich environment where learning algorithms can identify new terms and relationships to add to a model, remove outdated or irrelevant terms and relationships, and modify weights as certain paths become more popular.

Here we present a working prototype that allows system-guided search of document collections using automatically constructed domain knowledge as well as existing knowledge structures as used in digital libraries.

## 2 Related Work

There is a wealth of related work in log analysis, interactive search and other areas, e.g., [5, 12]. Due to limited space we will only present a few findings that should serve as motivations for our own work. First of all, we know that users are reluctant to leave any explicit feedback when they search a document collection [10]. However, implicit feedback, e.g., the analysis of log records, has been shown to be good at approximating explicit feedback. For example, users often reformulate their query and such patterns can help in learning an improved ranking function [6]. The same methods have shown to improve an adaptive domain model on a local Web site [8].

We can ask, however, do users want assisted search in the first place? First of all, digital libraries are characterized by much more structured knowledge than Web sites. This makes system-guided search a natural option as evidenced by the success of Aquabrowser[2] as a tool to access digital libraries. More generally though, there is also evidence that users want support in proposing keywords but they ultimately want to stay in control about what is being submitted as a query [16]. Furthermore, despite the risk of offering irrelevant suggestions in a system-guided search system, users might prefer having them rather than not [15]. On the other hand, it has also been shown that users are more inclined to submit new queries or resubmit modified queries than to navigate from the result set in a search environment that supports search *and* navigation [11]. Perhaps the best evidence for an interactive search system is the fact that all big Web search engines have recently added more and more interactive features, e.g., Google's Wonderwheel[3].

We are in line with what Belkin calls the *challenge of all challenges* in IR at the moment, to move beyond the limited, inherently non-interactive models of IR to truly interactive systems [2]. Our aim is to go beyond static interaction patterns and move to adaptive retrieval exploiting the implicit feedback that users leave when searching and navigating a document collection. Building adaptive domain models for digital libraries and other collections is our approach to capturing and utilizing collective intelligence [13].

[2] `http://www.aquabrowser.com/`
[3] `http://www.googlewonderwheel.com`

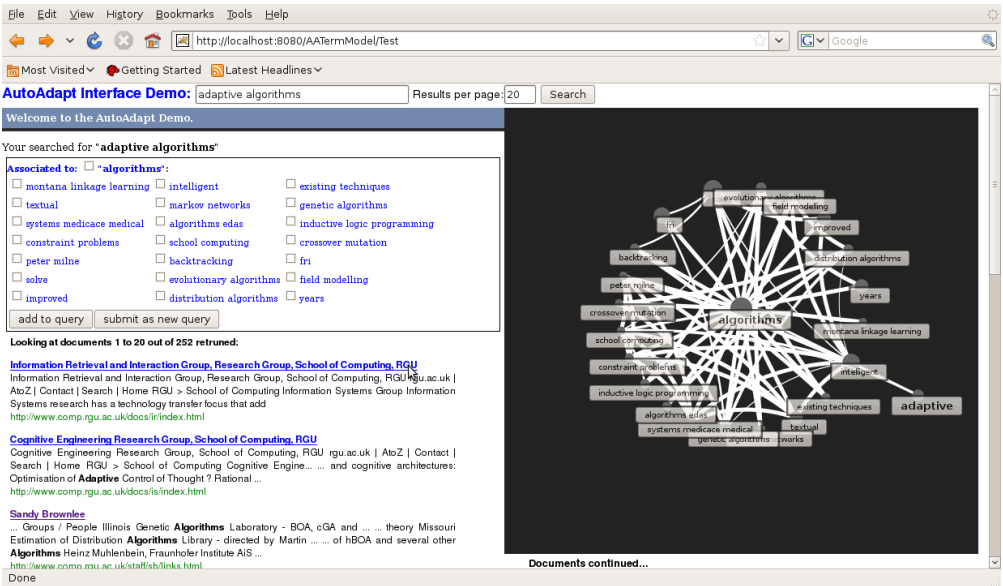



**Fig. 1.** Screenshot of AutoAdapt Demo System.

## 3 AutoAdapt Prototype

In Figure 1, we can see a screenshot of our demonstration system running on an intranet. In this particular case the domain model is automatically extracted from the document collection. The user submits a search query, this results in a number of matches (documents, book titles, etc.) being returned. Using the query terms, a segment of the domain model is displayed. The user can traverse the domain model by clicking on displayed terms. On term selection the list of suggested terms is updated. The user can then add the term to the existing query or use as a new query. The graph representations of domain models has been discussed in the literature, e.g. [4]. They are not the focus of our research but a useful tool.

The logging structure records a number of user decisions without the need for explicit feedback. What we are logging is not simply the action a user has taken (e.g., selecting a query modification, clicking a term in the domain model, selecting a match) but also recording what options a user has *not* taken but which have been available. This provides relative judgements that can be used to train classifiers [6, 8].

We have started to apply a number of techniques to turn the logged interactions into adaptive models which will be the focus of the next stage in the AutoAdapt project.

## Acknowledgements

AutoAdapt is funded by EPSRC grants EP/F035357/1 and EP/F035705/1. The JIT visualisation toolkit[4] was used for the domain model visualisation.

[4] `http://blog.thejit.org/javascript-information-visualization-toolkit-jit`

## References


1. M. J. Bates. The cascade of interactions in the digital library interface. *Information Processing and Management*, 38(3):381–400, 2002.
2. N. J. Belkin. Some(what) grand challenges for information retrieval. *SIGIR Forum*, 42(1):47–54, 2008.
3. S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, pages 107–117, Brisbane, 1998.
4. J. W. Buzydlowski, H. D. White, and X. Lin. Term co-occurrence analysis as an interface for digital libraries. In *Joint Conference on Digital Libraries*, pages 133–144, 2001.
5. J. Jansen, A. Spink, and I. Taksa, editors. *Handbook of Research on Web Log Analysis*. IGI, 2008.
6. T. Joachims and F. Radlinski. Search engines that learn from implicit feedback. *IEEE Computer*, 40(8):34–40, 2007.
7. J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the $9^{th}$ ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677. ACM, 1998.
8. D. Lungley and U. Kruschwitz. Automatically maintained domain knowledge: Initial findings. In *Proceedings of ECIR*, pages 739–743, 2009.
9. C. Manning, R. Prabhakar, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.
10. K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(8):1071–1081, June 2007.
11. M. Mat-Hassan and M. Levene. Associating Search and Navigation Behavior Through Log Analysis. *JASIST*, 56(9):913–934, 2005.
12. F. Silvestri. *Mining Query Logs: Turning Search Usage Data into Knowledge*. Foundations and Trends in Information Retrieval. Now Publisher, 2009. Forthcoming.
13. J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.
14. M. White. *Making Search Work: Implementing Web, Intranet and Enterprise Search*. Facet Publishing, 2007.
15. R. W. White, M. Bilenko, and S. Cucerzan. Studying the Use of Popular Destinations to Enhance Web Search Interaction. In *Proceedings of SIGIR'07*, pages 159–166, Amsterdam, 2007.
16. R. W. White and I. Ruthven. A Study of Interface Support Mechanisms for Interactive Information Retrieval. *JASIST*, 57(7):933–948, 2006.

# Providing multilingual subject access through linking of subject heading languages: The MACS approach

Patrice Landry[1]

[1] Swiss National Library, Hallwylstrasse 15, 3003 Bern, Switzerland

**Abstract.** The MACS project aims at providing multilingual subject access to library catalogues through the use of concordances between subject headings from LCSH, RAMEAU and SWD. The manual approach, as used by MACS, has been up to now the most reliable method for ensuring accurate multilingual subject access to bibliographic data. The presentation will give an overview on the development of the project and will outline the strategy and methods used by the MACS project. The presentation will also include a demonstration of the search interface developed by The European Library (TEL).

**Keywords:** MACS Project, multilingual subject access, LCSH, RAMEAU, SWD, mapping, concordances.

## 1 Introduction

Providing subject access using several languages in library online catalogues has become an important challenge for many national libraries in Europe. Given that the Web enables anyone to search any catalogue online, subject access using a single language becomes, for many national libraries in Europe, an important obstacle to efficient bibliographic retrieval. Aware of this obstacle and wishing to take advantage of the wealth of networks and the availability of rich and sustainable indexing languages, in 1998 four national libraries established the MACS project (Multilingual Access to Subjects)[1] under the auspices of the Conference of European National librarians (CENL).

Ten years later, the project has achieved this goal of providing the tools for multilingual subject access. It is now operational at the Swiss National Library and the Deutsche Nationalbibliothek where SWD (Schlagwortnormdatei) terms are added to links with RAMEAU, the French indexing subject headings list and LCSH (Library of Congress Subject Headings). The product is in the process of being integrated in The European Library portal [2].

## 2 MACS approach

Controlled vocabularies offer tremendous search possibilities for expanded use in web based services. Each vocabulary's authority file, for example subject heading lists, can contain upward of 300'000 to 500'000 entries. In controlled vocabularies, each term, name or subject, should have the same form each time it occurs in a bibliographic record or metadata. Relationships are established to ensure that accepted and rejected terms are linked and that users are directed to data identified under the preferred term. In recent years, there have been major developments in making these bibliographic metadata, such as terms (headings), their relationships and descriptions available as resources on the web. The issue is how these resources can be used in a meaningful and efficient way to improve information retrieval in the web environment. One of the solutions used in the context of subject headings has been the development of the SKOS (Simple Knowledge Organisation System) data model that provides a standard process for creating or migrating existing controlled vocabularies to the web.

While this solution offers the potential of being used in automatic matching or linking systems (ontology alignment), the MACS approach [3], [4], [5], [6], [7], [8] of linking terms manually from various subject heading languages offers the possibility of creating reliable and stable sets of linked data. Using as its base 90'000 RAMEAU-LCSH links, the project will gradually add SWD terms to these links and create new links with SWD headings. The manual method does not offer the same quantitative level results as automatic linking systems, but linking initiatives such as MACS create standard based sets of linked metadata that are relevant to large library collections.

The MACS linking methodology allow for links to be established based on an analysis at the terminology level (subject heading), at the semantic level (authority record), and at the syntactic level (indexing). A match or link is considered successful when a concept represented by similar headings in the different SHLs, which are matched manually (intellectually), returns the most closely equivalent results (titles) through subject retrieval. The MACS approach can be summarized by the following principles:

- Equality of languages and SHLs: No language or SHL is used as a pivot; each language is managed autonomously outside of MACS
- Establishment of equivalences between SHLs: Headings are not translated
- Equivalence links conceived as concept clusters: Mapping is done on the basis of concepts represented by SHLs
- Consistency of results: The quality of linking is based on the retrieval of consistent and similar sets of bibliographic records from different sources
- Extendable to other SHLs: Concept clusters can be increased by almost unlimited SHLs

Since March 2007, the Swiss National Library is creating links with SWD headings. This task was successfully integrated into the normal work process of the library and is now part of the workload of indexers. The Deutsche Nationalbibliothek has received funding to hire staff to work exclusively on link creation and work started in April 2009 at the Deutsche Nationalbibliothek in Leipzig. Their target is to

add an additional 45'000 links with SWD in the MACS database. By August 2009, close to 40'000 SWD had already been added to MACS links and it is planned that by 2010, most widely used topical headings used for indexing will have been covered. It is expected that 70'000 links with SWD will then be available and those links will cover 80% of bibliographic records of the Deutsche Nationalbibliothek's collections.

The MACS project has been involved in the TELplus project [9] that is investigating full text indexing and semantic search by providing access to its linking database. The work of TELplus in developing an automatic alignment of vocabularies using the semantic data of controlled vocabularies has provided MACS with a possible complementary linking strategy. The TELplus project has demonstrated that it can produce relatively reliable or relevant links in about 50% of the cases, mostly in the alignment of non ambiguous terms. Their method of evaluating these matches was to use MACS data as a comparison. The major challenge will be to test this methodology in searching different catalogues and finding equivalent results through subject retrieval. Up to now, manual subject headings linking projects has produced the most reliable linked data and it will be interesting to pursue alternate linking strategies.

## 3 Search Interface

With the growing amount of links in three languages, the task of developing a search interface was also on the MACS project's agenda. The strategy adopted by the MACS partners for realising this goal was to work with The European Library (TEL). Since TEL is a CENL service and that MACS partners are members of the CENL, it was considered to develop a search interface that could be integrated into the TEL portal. Discussions with the TEL office started in 2006 in the context of the EDL Project, one of whose objectives was the development of multilingual capacities of the TEL portal. A first prototype was created in October 2007 and work is presently underway to produce a new version of the search interface (LVAT II). The LVAT II project will address the issue of conducting "exact subject" queries across partner's library catalogues by building a central index using the open source indexing / search engine SOLR/LUCENE. Collections from the MACS partners will then be indexed in this central index. The project will also investigate the integration of TELplus[i] automatic subject heading alignments in the LVAT II.

Should that search interface project be successful, it will be left to CENL to formally integrate the MACS results in the TEL bibliographic services. This connection to TEL could lead to a permanent management solution to ensure the long term viability of the MACS initiative. The project is now confronted with demands to expand MACS to other languages (national and local) and would need to deal with the financial and management issues related to these demands. The Swiss National Library who has managed the project since its beginning has worked with the MACS partners to ensure funding for current development work. But with the potential of an increasing the number of participants, a new management structure would be needed correctly address these issues.

## References


1. http://macs.cenl.org
2. www.theeuropeanlibrary.org
3. Clavel-Merrin, G.: The Need for Co-Operation in Creating and Maintaining Multilingual Subject Authority Files", 65th IFLA Council and General Conference, Meeting 155. <http://WWW.ifla.org/IV/ifla65/papers/080-155e.htm> (1999)
4. Freyre, E., Naudi, M.: MACS: Subject Access Across Languages and Networks. In: Subject Retrieval in a Networked Environment: Papers Presented at an IFLA Satellite Meeting Sponsored by the IFLA Section on Classification and Indexing & IFLA Section on Information Technology, OCLC, *Dublin, Ohio, USA, 14-16 August 2001*. Dublin, OH: OCLC (2001)
5. MacEwan, A.: Crossing language Barriers in Europe: Linking LCSH to Other Subject Heading Languages. *Cataloging & Classification Quarterly* 29 (1/2), 199-207 (2000)
6. Landry, P.: The MACS Project: Multilingual Access to Subject (LCSH, RAMEAU, SWD). International Cataloguing and Bibliographic Control 30 (3), 46-49 (2001)
7. Landry, P.: Multilingual Subject Access: The Linking Approach of MACS. Cataloging & Classification Quarterly 31 (3-4), 177-191 (2004)
8. Landry, P.: The Evolution of Subject Heading Languages in Europe and their Impact on Subject Access Interoperability. In New Perspectives on Subject Indexing and Classification: Essays in Honour of Magda Heiner-Freiling, pp. 249-256. Deutsche Nationalbibliothek, Frankfurt am Main (2008)
9. www.theeuropeanlibrary.org/telplus
10. Wang, S., Issac, A., Schopman, B., Schlobach, S., Van der Meij, L.: Matching multilingual subject vocabularies (submitted for publication)

# Application Profiles Supporting Cross-Language and other Functionalities for Library Metadata

Barbara Levergood[1], Sally Chambers[2], Luigi Siciliano[3]

[1] Niedersächsische Staats- und Universitätsbibliothek Göttingen, Papendiek 14, 37073 Göttingen, Germany
levergood@sub.uni-goettingen.de
[2] The European Library, The National Library of the Netherlands, PO Box 90407, 2509 LK, The Hague, The Netherlands
Sally.Chambers@KB.nl
[3] University Library, Free University of Bozen-Bolzano, Universitätsplatz 1 - piazza Università, 1, 39100 Bozen-Bolzano, Italy
Luigi.Siciliano@unibz.it

**Abstract.** The CACAO project provides an infrastructure that enables cross-language functionality in digital libraries and library catalogues. The European Library is a free service that aggregates the bibliographic and digital collections of Europe's national libraries via a single multilingual interface. The involvement of The European Library in the CACAO project has assisted in the development of the CACAO Application Profile and has furthermore facilitated the development of The European Library Application Profile for Objects to better facilitate cross-language searching.

**Keywords:** CACAO Project, The European Library, application profiles, metadata, Dublin Core, digital libraries, library catalogues, cross-language, multilingual.

## 1 Introduction

The CACAO project[1] provides an infrastructure that enables cross-language functionality in digital libraries and library catalogues using CACAO's information retrieval and natural language processing (NLP) technologies. Through CACAO, the end-user can enter a query in his/her own language and retrieve documents and objects in any supported language. The European Library (TEL) is a free service that offers access to the bibliographic and digital collections of Europe's national libraries via a single multilingual interface.

This paper describes two application profiles (APs) for the metadata to be ingested by CACAO that have helped to facilitate aggregation and improve CACAO performance [1]. We discuss how the involvement of The European Library in

[1] CACAO Project (Cross-language Access to Catalogues and Online Libraries) is a 24-month targeted project supported by the eContentplus Programme of the European Commission. http://www.cacaoproject.eu/

CACAO has assisted in the development of one of the CACAO APs based on The European Library Application Profile for Objects (TEL-AP for Objects) [2].

## 2 CACAO's Application Profiles

As defined by Heery and Patel, application profiles are "schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application" [3]. The CACAO AP working group identified several requirements that influence the choice of metadata formats. Since CACAO wanted to harvest metadata from OAI-PMH [4] repositories, the Simple Dublin Core [5] required by the OAI-PMH Guidelines [6] should be among the formats selected by CACAO. The metadata should be in a format that could be reused. The format and encoding of the metadata should be mature, readily available, and easy to implement. The CACAO AP should be based on an AP that CACAO would have to implement anyway. The AP would need to be flexible, offering a sensible requirement for a minimum record, i.e. title and identifier, but also rich metadata supporting CACAO's NLP-based cross-language services, with language specifications, subject headings, classification notations, alternative titles, table of contents, etc. CACAO's solution was to create two APs, one based on Simple Dublin Core and one based on the TEL-AP for Objects, which is in turn based on the Dublin Core Library Application Profile (DC-Lib) [7], itself Qualified Dublin Core-based.

The next task was to identify the requirements relevant at the element- and attribute-levels. First, CACAO needs to analyze the text of certain fields as a part of the indexing process. The language of the metadata is used if available, otherwise the language of the resource or a language guesser may be used. The predominant language of a vocabulary encoding scheme (VES) used in a subject field might also be used as an imperfect substitute. Second, CACAO is experimenting with a word sense disambiguation tool, Word2Category [8], [9], that associates words with classifications in a classification system such as Dewey Decimal Classification [10]. In order to perform this association, the language of the metadata and the classification system must be identified. Third, interfaces for cross-language services need to support those services; searching, facets and the display may all draw on information about the language of the resource.

Based on these requirements, we offer some best practices for how APs might be optimized for cross-language functionalities. All can be implemented in Qualified Dublin Core. Only the identification of the VES is not possible in Simple Dublin Core. All are deviations from the TEL-AP for Objects. (1) Recommend the use of `dc:language` for the language of the resource. (2) Recommend the use of `xml:lang` for the language of the metadata for text elements that contain semantically important content.[2] (3) Recommend the use of `xsi:type` for the

[2] Note, however, that one of the Qualified DC XML Schemas, version 2008-02-11, http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd, prohibits the use of the `xml:lang` attribute with a number of vocabulary encoding schemes, including LCSH.

identification of the VESs of subject fields. (4) Provide an XML schema which permits local customizations such as the identification of the VESs of subject fields via attributes or the addition of new elements.

## 3 The European Library Application Profile for Objects

In order for The European Library to aggregate the collections from Europe's national libraries, an interoperable metadata format was needed. Within the national library community, it was felt that "TEL will use a collection of namespaces, among which the DC-Lib will be the most important, although this will probably not be sufficient for TEL" [11]. The decision was therefore made that the TEL project[3] would develop its own AP with DC-Lib as the basis.

Three remaining key concerns were identified. (1) The need for new collection-level metadata fields in order to provide collection-level services led to the development of The European Library Application Profile for Collection Descriptions [12]. (2) The need for an identifier for retrieving the metadata record from its originating collection was solved by introducing a new term in the TEL namespace, `tel:recordId`. (3) The need for a link to the digital object to permit direct access was addressed by using the existing Dublin Core term `dc:identifier` in combination with internal coding in the portal itself.

Since the original specification, new requirements have been identified that have necessitated the addition of new terms in the TEL namespace. For instance, as part of the TELplus project [13], 20 million pages of OCR'd material will be made available in The European Library. Partner libraries will provide links in their metadata records to this full-text for indexing purposes, requiring the addition of new metadata elements in v2.0 of the TEL-AP for Objects.

The project Europeana v1.0 [14] will develop the Europeana prototype [15], launched in November 2008, into a full operational service. It is anticipated that The European Library will then become the domain-level aggregator for libraries in Europeana. With this in mind, it is intended that the TEL-AP for Objects will evolve further to ensure that it is interoperable with the Europeana metadata format, European Semantic Elements (ESE) [16]. In addition, any new requirements from the wider library community, such as university and research libraries, expected to be the first group of libraries to join the national libraries in The European Library, will be taken into consideration. In addition, The European Library is actively involved in the Accessible Registries of Rights Information and Orphan Works towards Europeana (ARROW) project [17], whose goal is to develop a digital rights infrastructure for Europe. It is anticipated that the TEL-AP for Objects either will be used as an intermediary within the proposed digital rights infrastructure, or will be extended to

[3] The European Library (TEL) Project was a 30 month project beginning in 2001, funded under the European Commission's 5th Framework Programme, http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive%5Ctelproject_archive/telproject_archive.html.

provide a link to rights information held elsewhere in a dedicated rights metadata format such as in the ONIX framework.[4]

**Acknowledgments.** The authors wish to thank Raffaella Bernardi, Alessio Bosca, Paolo Buoso, Stefan Farrenkopf, Daniele Gobbetti, Stefanie Rühle, Romain Wenz and each other for the stimulating discussions that led to the development of the CACAO APs.

## References

1. Levergood, B., Siciliano, L., Gobbetti, D., Dini, L., Bosca, A., Buoso, P., Barsanti, I.: Integration with www.theeuropeanlibrary.org and aggregation of partner libraries. CACAO D5.2 (public) (2009)
2. The European Library Metadata Registry (for objects), http://www.theeuropeanlibrary.org/handbook/regtable.php
3. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. Ariadne 25 (2000), http://www.ariadne.ac.uk/issue25/app-profiles/
4. Open Archives Initiative - Protocol for Metadata Harvesting, http://www.openarchives.org/pmh/
5 Dublin Core Metadata Initiative, http://dublincore.org/
6. Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers, http://www.openarchives.org/OAI/2.0/guidelines-repository.htm
7. Dublin Core Libraries Application Profile, April 2002, http://dublincore.org/documents/2002/04/16/library-application-profile/, was used as the basis.
8. Bosca, A., Gobbetti, D.: Fully integrated CLIR system. CACAO D.1.4 (confidential) (2008)
9. Levergood, B., Farrenkopf, S., Frasnelli, E.: The Specification of the Language of the Field and Interoperability: Cross-language Access to Catalogues and Online Libraries (CACAO). In: Greenberg, J., Klas, W. (eds.) Metadata for Semantic and Social Applications: Proceedings of the International Conference on Dublin Core and Metadata Applications 22-26 September 2008, pp. 191-196, Universitätsverlag, Göttingen (2008), http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf
10. Dewey Decimal Classification, http://www.oclc.org/dewey/
11. Minutes of the TELproject (WP3: Metadata Development) meeting, 1 February 2002. (Unpublished)
12. The European Library Application Profile for Collection Descriptions (v1.5), http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap_cld.html
13. TELplus project, http://www.theeuropeanlibrary.org/telplus
14. Europeana v1.0, http://version1.europeana.eu/
15. Europeana, http://www.europeana.eu/portal/
16. Specification for the Europeana Semantic Elements (v3.2), http://www.version1.europeana.eu/web/guest/provide_content
17. ARROW, Accessible Registries of Rights Information and Orphan Works towards Europeana, http://www.arrow-net.eu/

[4] The ONIX family includes standards for Books, Serials and Licensing Terms, http://www.editeur.org/8/ONIX/

# Content Extraction Meets the Social Web in the LiveMemories Project

Massimo Poesio
University of Trento

Bernardo Magnini
FBK-irst, Trento, Italy

## Introduction

The widespread availability of low-cost means for putting into digital form multimodal information including text of both traditional and non-traditional type (from stories to blogs), photos, and videos, together with the explosion in use of social networking sites such as Facebook for publishing such information about oneself or about the events of the day on the Web and sharing it with friends and perfect strangers, is leading to radically new forms for the preservation of information and the creation of collective memory.

However, the functionality offered by current social networking sites does not go beyond the upload and indexing of such information; indexing is most often word-based or at most topic-based, and the data thus collected lie otherwise unanalyzed. The aim of LiveMemories[1] is to take advantage of techniques for extracting content from multimedia sources to make such shared digital repositories 'alive' by identifying people and objects mentioned in them, and extracting information about events and other relations between objects, including temporal information.

This type of analysis will enable new presentation methods. Consider the following scenario. Luisa Tomasi from Gardolo goes to a Franco Battiato concert in Trento on February 17th and 18th, and takes some pictures. The next day she creates a description of the event on the LiveMemories portal, uploading the images she took and accompanying them with text describing her experience and giving her comments on the concert. Images and text are analyzed by the LiveMemories platform, that recognizes the event as one listed on www.crushsite.it and identifies Franco Battiato as one of the individuals stored in its knowledge base (automatically extracted by processing WikiPedia text and text from the local press) indeed. LiveMemories can then offer to Luisa further information about the event, e.g., it can tell Luisa that the brilliant viola player is called Demetrio Comuzzi, or it may offer to Luisa to visualize Battiato's discography in the form of a chronology to discover when a particular song came out. LiveMemories may also discover that other people with an account on the portal went to that same concert including e.g. Mario Boato, who also uploaded his

[1] LiveMemories is a three years project funded by the Autonomous Province of Trento. The project, started in October 2008, is a collaboration among three academic partners, Fondazione Bruno Kessler (FBK), University of Trento (Italy), and University of Southampton (UK), and a number of companies and data providers located in the Trentino area. Detailed information can be found at the project web site: http://www.livememories.org.

own data. LiveMemories can point this out to Luisa and Mario, who may also discover they share a preference for Battiato's early music.

LiveMemories includes activities in Content Extraction from text and images, Content Presentation, and Content Integration. In this paper, we will focus on Content Extraction from text and on cross-document coreference.

## Background

The availability of high-performance tools for POS tagging and parsing has made it possible to contemplate large-scale semantic processing (named entity extraction, coreference, relation extraction, ontology population). US initiatives such as MUC, ACE and GALE made large annotated resources available and introduced quantitative evaluation.

In intra-document coreference (IDC) this led to the development of the first large-scale machine learning models using these resources and to the development of IDC tools, most recently, the ELKFED/BART system (Versley et al., 2008). In relation extraction, work carried out as part of the ACE initiative and in ELERFED showed that good results can be obtained extracting relations from news with supervised methods, particularly Support Vector Machines (SVMs) and Kernel Methods but that semi-supervised methods are more effective with less formal text.

Interest in cross-document coreference (CDC) has began fairly recently (Bagga and Baldwin, 1998), but there has been much development in recent years because of great interest both from government and from industry. In particular there has been great interest in a simpler form of entity disambiguation, generally known as Web entity as in the case of the Web people task of Semeval (Popescu and Magnini, 2007) and the Spock challenge[2]. As testified by the SEMEVAL Web People task[3] , most state of the art systems are based on clustering of entity descriptions containing a mixture of collocational and other information, among which information about entities and relations. SEMEVAL also showed that the clustering technique and especially the termination criterion are crucial. Finally, work on the Spock challenge highlighted the need for methods for handling huge quantities of information. Recent developments have therefore focused on improving the clustering technique and experimenting with different types of information that can be extracted robustly from text. (See, e.g., the results with ELERFED.) Most of the work discussed above was carried out for English; progress with languages other than English includes work on German (e.g., Versley) and Spanish (Ferrandez) but very little on Italian apart from work by Delmonte, also in part for lack of resources. In this direction it is worth to mention the creation of a reference benchmark for CDC in Italian (see Bentivogli et al. (2008)), which is used and improved in the LiveMemories project.

[2] challenge.spock.com/.
[3] http://nlp.uned.es/weps/.

# Goals of LiveMemories: Content Extraction from Text

LiveMemories builds on top of previous content extraction technologies. Particularly, we use TextPro[4] (Pianta et al. (2008)) a suite of modular Natural Language Processing tools for analysis of Italian and English texts. The current version of the tool suite provides functions ranging from tokenization to chunking and Named Entity Recognition. TextPro performed the best on the task of Italian NER and Italian PoS Tagging at EVALITA 2007[5].

The performance and usefulness of existing technology for content extraction from text is currently improved by:

- Larger corpora and techniques, obviating the need for large scale annotation (e.g., active learning, weakly supervised methods).
- Better preprocessing techniques (often underestimated);
- Incorporating automatically extracted lexical and commonsense features in addition to traditional 'surface' features;
- Developing better Machine Learning methods to exploit these more advanced sources of information (e.g. kernel functions)
- Developing richer representations of relations, e.g., with temporal modification (e.g. *John Doe was CFO of ACME from 2001 to 2005*);
- Further developing automatic methods for Textual Entailment Recognition, a robust type of textual inference based on patterns that can be automatically acquired from corpora. We use the EDITS system[6] (Negri et al. (2009)).

# Conclusions

In the first year of the project, we made substantial progress in content extraction, including:

- The development of a new cross-document coreference resolver based on the work by Popescu (Popescu and Magnini, 2007);
- The creation of a new annotated corpus for coreference in Italian;
- The development of a new intra-document coreference resolver for Italian based on BART;

In addition, we made lot of progress on building a user community, establishing contact with a number of communities in Trentino that will become pilot users of our platform; and designed a new platform for digital memories centered around the notion of **story**. The first release of the platform will become available end of September 2009.

[4]TextPro is freely distributed for research purposes at http://textpro.fbk.eu/.
[5]http://evalita.fbk.eu/.
[6]EDITS is distributed as open source software at http://edits.fbk.eu/.

# References

Bagga, A. and Baldwin, B.(1988), *Entity-Based Cross-Document Coreferencing Using the Vector Space Model*, in *Proceedings of COLING/ACL*, Montreal, 1998.

Bentivogli L., Girardi C. and Pianta E. (2008), *Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News*, in *Proceedings of LREC 2008 Workshop on Resources and Evaluation for Identity Matching, Entity Resolution and Entity Management*, Marrakech, Morocco.

Negri M., Kouylekov M., Magnini B., Mehdad Y. and Cabrio E. (2009) *Towards Extensible Textual Entailment Engines: the EDITS Package*, in *Proceedings of the XI Conference of the Italian Association for Artificial Intelligence - to appear*, Reggio Emilia, Italy.

Pianta E., Girardi C. and Zanoli R. (2008), *The TextPro tool suite*, in *Proceedings of LREC, 6th edition of the Language Resources and Evaluation Conference*, Marrakech, Morocco.

Popescu, O. and Magnini, B. (2007) *IRST-BP: Web People Search Using Name Entities*, in *Proceedings of SEMEVAL*, 2007.

Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X. and Moschitti, A. (2008), *BART: A modular toolkit for coreference resolution*, in *Proceedings of LREC*, Marrakesh, 2008.

# Tools for Document Image Retrieval in Digital Libraries: the AIDI System

Simone Marinai, Giovanni Soda

Dipartimento di Sistemi e Informatica
University of Florence, Italy

In the last few years, Digital Libraries became one important application area for Document Image Analysis and Recognition research [1]. In this field, a relevant line of research is Document Image Retrieval (DIR) that aims at finding relevant documents relying on image features only. DIR techniques are used to index not only the textual content of a document, but also its layout, graphical objects, mathematical equations, and handwritten text. By integrating these capabilities with other traditional indexing approaches we expect it will be possible to define new search strategies that could be especially useful in scientific and technical collections. In this abstract we summarize our recent research on Document Image Retrieval techniques in the field of Digital Libraries that we integrated in the AIDI prototype system.

## 1 Document Image Retrieval techniques

Document Image Retrieval aims at finding relevant documents from a corpus of digitized pages relying on image features only and is closely related to Content-Based Image Retrieval (CBIR) [2] [3]. Important sub-tasks include the retrieval of documents on the basis of layout similarity and on the basis of the textual content.

### 1.1 Word indexing

One very important sub-topic of text-based DIR is word-level indexing, that addresses the efficient identification of the occurrences of a given word in the indexed documents (e.g. [4] [5] [6] [7]). When the use of OCR is not advisable, either due to the low quality of images or the use of uncommon fonts, then image-based word retrieval is a viable alternative. In methods based on character-like coding some objects (that might correspond to characters) are extracted from each word. The word is then represented by concatenating the codes assigned to the objects on the basis of shape similarity [5]. The word indexing implemented in the AIDI system relies on character-like coding and is described with more details in [6].

### 1.2 Layout Indexing

The layout of a page conveys some semantics that is important for both scholars and general readers, but is often neglected by DL's indexing approaches. For instance, users could be interested on identifying the pages having a *marginalia*

in the right side, or would like to retrieve a page containing a figure on some specific position in the left column in the page. Another example is the retrieval of the title page of scientific papers that, starting from a general structure, can have different actual layouts. Some systems have been proposed in the past to index various types of documents, such as forms [8] [9] and journal papers [10].

In the AIDI system we represented the page layout with a hierarchical description based on the XY tree [11]. XY trees have been demonstrated to be useful when dealing with documents containing ruling lines and can deal with multi-column pages as well as with pages where the pictures cover more text columns [10]. Leaves of the tree correspond to homogeneous regions in the page. To perform the page retrieval, the MXY trees are encoded into a fixed-size representation that is subsequently used to rank the pages. The layout indexing implemented in the AIDI prototype has been described in various papers (e.g. see [10] [12]).

### 1.3 Early printed books

Early printed books, such as the Latin Gutenberg Bible, look very similar to medieval manuscripts, since they contain illuminated letters (hand painted) and several ligatures and abbreviations that were standard in manuscript writing and have been slowly abandoned in the technological progress of printing. Indexing early printed documents is therefore a task that is closely related to handwriting indexing. To demonstrate the feasibility of these approaches, we recently addressed the indexing and retrieval of the Gutenberg Bible with a Query by Example (QbE) retrieval mechanism implemented in a prototype tool [13].

### 1.4 Mathematical Symbol Indexing

The recognition of mathematical symbols is particularly difficult for three main reasons: the very large number of symbol classes, the reduced script size for superscripts and subscripts, and the lack of linguistic tools (such as dictionaries) that could help in the recognition. Document image retrieval techniques have been seldom used to process mathematical expressions. However, several researchers envisage the utility of math search systems that would be able not only to search for text, but also for "fine-grain mathematical data" (e.g. equations and functions) [14]. In [15] we recently described our current work on the development of one mathematical symbol indexing and retrieval module.

## 2 The AIDI system

The Automatic Indexing of Document Images (AIDI) system, that has been developed by our research group, integrates a font-independent word indexing and a layout-based document retrieval into a unique framework [10] [16]. Figure 1 shows a snapshot of the AIDI user interface. On the top-left there are ten thumbnails that contain either the browsed pages or the retrieval results. The image on the right is one selected page that can be further enlarged in the zoom area. The bottom-left part contains the buttons used to perform the queries. In

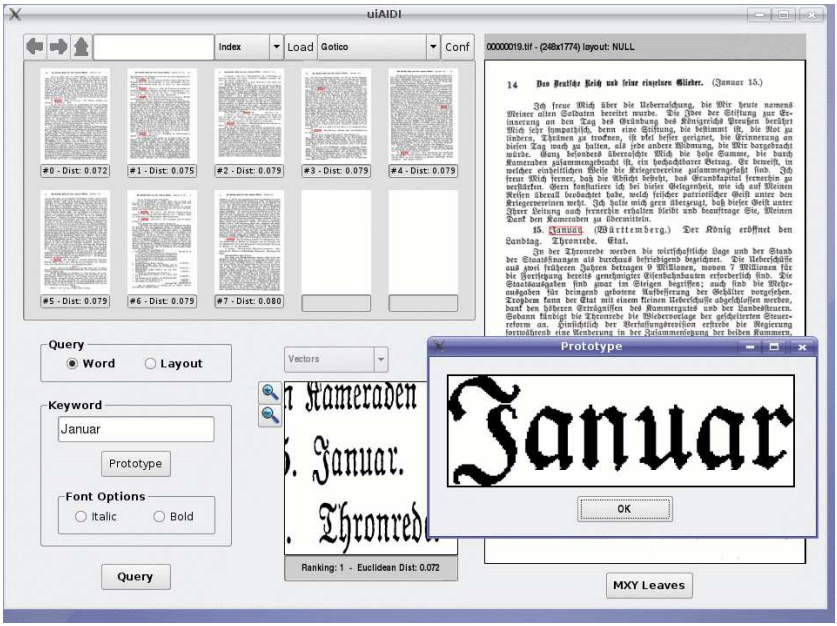



**Fig. 1.** The user interface of the AIDI system.

the case of textual queries the user enters the query word in the appropriate field. The "Prototype" window shows the generated prototype (in this case a word printed with the Gothic font). Layout-based queries are made with a QbE approach. Therefore, the user selects a page of interest from the list of thumbnails and performs the query by pressing the appropriate button.

During the indexing, the pages are first processed by a layout analysis tool that extracts homogeneous regions. Textual regions are subsequently analyzed so as to identify the words, that are encoded with appropriate character labels. At the same time the layout is encoded to obtain a page-level representation of the documents. The pages can be retrieved by taking into account both textual and layout queries. In the first case, a query prototype is obtained by rendering the word entered by the user with the LaTeX package (see the "Prototype" window in the Figure). The prototype is encoded similarly to the indexed words that are lastly ranked according to their similarity with the query. Likewise, a query page can be represented in the same way of indexed pages that can be ranked according to their layout similarity.

## 3 Conclusions

Image-based indexing techniques can be adopted for large collections only if scalable approaches are available to index feature vectors. We are currently testing some indexing techniques to the problem of word image indexing with interesting results for both effectiveness and efficiency of the retrieval [17].

## References


1. Baird, H.S.: Digital libraries and document image analysis. In: Int'l Conference on Document Analysis and Recognition. (2003) 2–14
2. Doermann, D.: The indexing and retrieval of document images: A survey. Computer Vision and Image Understanding **70**(3) (June 1998) 287–298
3. Mitra, M., Chaudhuri, B.: Information retrieval from documents: A survey. Information Retrieval **2**(2/3) (2000) 141–163
4. Rath, T.M., Manmatha, R.: Word spotting for historical documents. IJDAR **9**(2-4) (2007) 139–152
5. Lu, S., Li, L., Tan, C.L.: Document image retrieval through word shape coding. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11) (Nov. 2008) 1913–1918
6. Marinai, S., Marino, E., Soda, G.: Font adaptive word indexing of modern printed documents. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(8) (2006) 1187–1199
7. Meshesha, M., Jawahar, C.V.: Matching word images for content-based retrieval from printed document images. IJDAR **11**(1) (2008) 29–38
8. Liu, J., Jain, A.: Image-based form document retrieval. Pattern Recognition **33** (2000) 503–513
9. Duygulu, P., Atalay, V.: A hierarchical representation of form documents for identification and retrieval. IJDAR **5**(1) (November 2002) 17–27
10. Marinai, S., Marino, E., Cesarini, F., Soda, G.: A general system for the retrieval of document images from digital libraries. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), 23-24 January 2004, Palo Alto, CA, USA, IEEE Computer Society (2004) 150–173
11. Nagy, G., Seth, S.: Hierarchical representation of optically scanned documents. In: Int'l Conference on Pattern Recognition. (1984) 347–349
12. Marinai, S., Marino, E., Soda, G.: Tree clustering for layout-based document image retrieval. In: Proc. 2nd International Workshop on Document Image Analysis for Libraries (DIAL 2006), 27-28 April 2006, Lyon France, IEEE Computer Society. (2006) 243–251
13. Marinai, S.: Text retrieval from early printed books. In: Third Workshop on Analytics for Noisy Unstructured Text Data, ACM Press (2009) 33–40
14. Youssef, A.: Roles of math search in mathematics. In: Mathematical Knowledge Management MKM 2006, Springer Verlag- LNCS 4108 (2006) 2–16
15. Marinai, S., Miotti, B., Soda, G.: Mathematical symbol indexing using topologically ordered clusters of shape context. In: Proc. 10th Int'l Conference on Document Analysis and Recognition, Washington, DC, USA, IEEE Computer Society (2009) 1041–1045
16. Marinai, S., Marino, E., Soda, G.: Exploring digital libraries with document image retrieval. In Kovács, L., Fuhr, N., Meghini, C., eds.: ECDL. Volume 4675 of Lecture Notes in Computer Science., Springer (2007) 368–379
17. Marinai, S., Marino, E., Soda, G.: Embedded map projection for dimensionality reduction-based similarity search. In: Proc. Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, SSPR/SPR. (2008) 582–591

# Creating and Aligning Controlled Vocabularies

Ahsan-ul Morshed
morshed@dit.unitn.com
Margherita Sini
margherita.sini@fao.org

[1] Department of Information and Communication Technology
University of Trento, Italy
[2] Food and Agriculture Organization of the United Nations (FAO)
Rome, Italy

**Abstract.** A vocabulary stores words, synonyms, word sense definitions (i.e. glosses), relations between word senses and concepts; such a vocabulary is generally referred to as the Controlled Vocabulary if choice or selections of terms are done by domain specialists. In our case,we create and match two controlled vocabularies by using their concept facets. This methodology is based on semantic matching which is different from the orthodox view of matching.

**Key words:** Vocabulary Mapping, Vocabulary Creation, Thesaurus, AGROVOC, CABI

## 1 Automatic Controlled Vocabulary Creation

Some research has been done on Controlled Vocabulary (CV) construction by automatic or semi-automatic methods [3]. These two methods can be categorized into two approaches [1]: In the **statistical approach**, terms are extracted from a document by IDF (inverse document frequency). Adapted to the controlled vocabulary construction problem, the assumption is that frequently co-occurring words with a text window (sentence, paragraph or whole text) point to some semantic cohesiveness. The co-occurrence approach needs human intervention before terms can be used for controlled vocabulary creations. From a **linguistic approach**, terms and their relations are based on the distributional context of syntactic unit (subject and object) and the grammatical surrounding function these unit. For example, suppose we have two terms "Agricultural business" and "Agricultural industry". These two terms can be semantically mapped:

- The above word terms shared the same head or tail (i.e. agricultural).
- The substituted words have the same grammatical function (Modifier, i.e. business and industry).
- The substituted words are semantically close (i.e. business and industry).

The two described approaches are time-consuming and need a substantial amount of human intervention. To overcome this problem, we combine the previously

cited two approaches into one. Furthermore, we have used semantic matching algorithm to find the relations among terms, reducing time compared to the linguistic techniques. Our approach is different from others because they use syntactic matching techniques and they do not make use of background knowledge. Because it is difficult to find the universal background knowledge, we used WordNet [7] in order to conduct testing.

Our algorithm is defined into micro steps as follows:

*Step 1:* Extracting terms from a document using NLP tools.

*Step 2:* Building Semantic Relationships among terms and using S-match tools [2] for calculating relatedness among the terms.

*Step 3:* Filtering Terms Relationships with WordNet/External Resources.

*Step 4:* Giving linkage information for words according to semantic similarities.

In Step 1 we take a set of documents and extract keywords using the Kea tool [5]. In Step 2 we use the Element Level Matcher from S-Match tool to calculate the relatedness between two terms. In Step 3 we use WordNet to filter the information. After filtering, we cluster keywords according to semantic similarities. This work on automatic CV creation is still on going: we have presented the general idea and described the algorithm, but more work would need to be carried out in order to extend the testing.

## 2 Controlled Vocabulary Matching

A Concept Facet (CF) contains distinct features for each concept: it includes combined relations, CF= $\langle lg, mg, R \rangle$, where $lg$ identifies less general concepts (one or more), $mg$ identifies more general concepts (one or more) and $R$ identifies related concepts (one or more). In order to realize a matching between two vocabularies (CV1, CV2), we consider the CF from all given CVs's concepts: for every CF of CV1, we check the matching with all CFs of CV2. These concept facets are stored in tables for matching purpose. The methodology of the matching algorithm applied to every concept, can be represented with the following picture. The matching between two concept facets follows the top-down

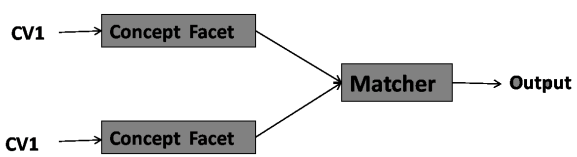



**Fig. 1.** CV Matching

approach and used several lexical comparison algorithms (SMOADistance, HammingDistance, JaroMeasure, SubStringDistance, N-gram, JaroWinKlerMeasure,

and LavesteinDistance) [4, 8]. Firstly, we start comparing the more general concepts; if they match (they have same lexicalizations or they are synonyms) we assume that the concepts under investigation belongs to same concept (they match). Secondly (either we got match or not), we start comparing the less general concepts. Based on the results of two mentioned matching, we may obtain exact match (in case more general and less general concepts match), partial match (in case of only one match), or not match. Related concepts of CFs are considered to validate the previous results.

## 3 Results and Evaluation: the AGROVOC and CABI case study

In our experiments, we used the AGROVOC thesaurus and the CABI thesaurus because there is no complete mapping between them. The results of the mapping will be published online so that users can use them for better indexing, searching and information retrieval [6, 11].

### 3.1 AGROVOC

AGROVOC is a multilingual controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. the environment). The AGROVOC Thesaurus was developed by FAO and the Commission of the European Communities in the early 1980s. Since then it has been updated continuously by FAO and local institutions in member countries. It is mainly used for indexing and retrieval data in agriculture information systems both inside and outside FAO. It has approximately 20,000 concepts and four types of relations derived from the ISO standard. Among the available format, we used the XML version for our task [9].

### 3.2 CABI

CABI is a monolingual controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, horticulture, soil science, entomology, mycology, parasitology, veterinary medicine, nutrition and rural studies. The CABI thesaurus was developed by CABI which is a not-for-profit, science-based development and information organization. It has 48,000 concepts and four types of relationship derived from the ISO standard. We obtained data as text format and converted it to XML format for experiment purposes [10].

### 3.3 Results and Evaluation Descriptions

We started our experiments using 492 concepts from each controlled vocabulary. Managing all concepts was a challenge because the two vocabularies are not organized in the same structure. We converted each vocabulary to the same format

in order to conduct the test. We obtained 64 exact matches from all tested algorithms, but we found different numbers of partial matches from eight element label matchers. SMOADistance matcher gives more partial matches than others. Hamming distance, JaroMeasure, SubStringDistance, and N-gram do not give a satisfactory numbers of matches. JaroWinKlerMesaure and LevesteinDistance produce quite similar results. However, these are our primary results which should be validated by extending the process to the full thesauri.

## 4 Conclusion

In this paper, we have shown our proposed system for automatic creation of controlled vocabulary and vocabulary matching using concept facets. We are convinced that it helps for better information searching, browsing, and extraction in agriculture and related domains. There are some open research issues: the semantic heterogeneity between two controlled vocabularies in a single domain; the multi-word concepts; the possibility of automatically link non-matched concepts to external reliable resources such as public thesauri, encyclopedia or dictionaries.

## Acknowledgment

Authors would like to thank Prof. Fausto Giunchiglia, Ilya Zaihrayeu, and Vincenzo Maltese for their valuable suggestions. Also, we would like to thank Shaun Hobbs of CABI for kindly providing us with the data files.

## References

1. F.Ibekwe-SanJuan. Construction and maintaining knowledge organization tools a symbolic appraoch. volume 62, 2006.
2. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-match: An algorithm and an implementation of semantic matching. *In Proceedings of ESWS'04*, 2004.
3. A.Gilchrist J.Aitchison and Bawden. Thesaurus construction and use:a practical manual. 4th ed., page 240, London, 2006. Aslib.
4. J.Euzenate and P.Shaviko. *Ontology Matching*. Springer, 1st edition, 2007.
5. KEA Automatic keyphrase extraction. http://www.nzdl.org/Kea/.
6. Sini M. Chang C. Li S. Lu W. He C. Liang, A. and J. Keizer. The mapping schema from chinese agricultural thesaurus to agrovoc. In Proceedings of the fifth Conference of the European Federation for Information Technology in Agriculture, Food and Environment and the thirdWorld Congress on Computers in Agriculture and Natural Resources, 2005.
7. George Miller. *WordNet: An electronic Lexical Database*. MIT Press, 1998.
8. Natalya F. Noy. Semantic integration: a survey of ontology-based approaches. *SIGMOD Rec.*, 33(4):65–70, 2004.
9. Agrovoc thesaurus. http://www.fao.org/agrovoc/.
10. CAB thesaurus. http://www.cabi.org/.
11. L.Finch H. Kolb W.Hage, M.Sini and G.Schreiber. The oaei food task:an analysis of a thesaurus alignment task.

# Personalization based on users' requirements in MANUSCRIPTORIUM (European Digital Library of Manuscripts)

Tomáš Psohlavec[1], Jakub Heller[2], Roberto Caldelli[3], Tomasz Parkoła[4]

[1] AiP Beroun s.r.o., Talichova 807, Beroun, 266 01, Czech Republic
[2] Cross Czech a.s., Růžová 17, Praha 1, 11000, Czech Republic
[3] Media Integration and Communication Centre Firenze, Viale Morgagni 65, 50134 Florence, Italy
[4] Poznan Supercomputing and Networking Center, Noskowskiego 12/14, 61704 Poznan, Poland
tp@aipberoun.cz; hellerj@crossczech.cz; caldelli@lci.det.unifi.it, tparkola@man.poznan.pl

**Abstract:** Manuscriptorium (http://www.manuscriptorium.eu) is a resource provided by the National Library of the Czech Republic (http://www.nkp.cz) as a strategic leader and content coordinator as well as by the AIP Beroun Ltd. (http://www.aipeberoun.cz) as a technical provider and system administrator. Manuscriptorium became a major resource at the European level due to realization of the ENRICH project (http://enrich.manuscriptorium.com) funded under eContent+ programme. The main results of the project are now available and therefore a short Manuscriptorium case study can be presented, including demonstration of new end-user features such as personalized collections and virtual documents.

**Keywords: Manuscripts, Incunabula, Digital Library, Interoperability, Shared repositories, User requirements**

## Introduction

The aim of the ENRICH project is the creation of a base for the European digital library research environment for study of specific historical cultural heritage consisting of manuscripts, incunabula, early printed books, historical archival materials, etc. The main innovation of ENRICH lies in a common easy-to-use interface which enables concentration of dispersed resources into a unique research environment and retrieval of data from distant servers. The project allows the users to search and access documents which would otherwise be hardly accessible by providing free access to almost all digitized manuscripts in Europe. During the demo session we will present selected partners documents/collections within the end-users interface in order to demonstrate aggregation results and at the same moment we will demonstrate the Manuscriptorium end-users features with special focus on the newly available "Personalized Virtual Library" feature. Four typical ways of cooperation will be shortly mentioned using the real-life examples.

## Manuscriptorium Platform

ENRICH builds upon the existing Manuscriptorium platform (http://www.manuscriptorium.com) adapted to needs of organizations holding repositories of manuscripts. Manuscriptorium is a resource provided on-line by the National Library of the Czech Republic (http://www.nkp.cz) as a coordinator as well as by the AiP Beroun Ltd. (http://www.aipeberoun.cz) as a technical provider and system administrator. A searchable Open catalogue of historical documents is an important part of the service along with the Digital Library that contains all digital documents aggregated so far. Various front-end interfaces are implemented into the system including the OAI-PMH and Z39.50 interfaces which ensures inclusion to partners portals (Manuscriptorium DL is searchable via OAI-PMH from the TEL portal, i.e. any ENRICH partner contributing to Manuscriptorium automatically enriches the European Digital Library) as well as OAI-PMH harvester which aggregates partner's documents into Manuscriptorium, set of convertors for performing various metadata formats conversions etc.

## Data Aggregation

The ENRICH groups together the richest owners of digitized manuscripts among national libraries in Europe; ENRICH partner libraries possess almost 85% currently digitized manuscripts in the national libraries in Europe, which is enhanced by substantial amount of data from university libraries and other types of institutions. The consortium will make available more than 5 076 000 of digitized pages by the end of November 2009.

The principle of integration is centralization of metadata within the Manuscriptorium digital library and distribution of data among other resources within the virtual net environment. The project creates conditions that enable the partners bringing together appropriate mass of digital content. Depending on level of readiness of a particular partner following general ways of cooperation are possible:

- For Advanced digitization projects operating digital libraries equipped with the OAI-PMH interface: partners provide Manuscriptorium with descriptive and structural metadata via one of their available OAI profiles. Metadata are harvested and processed within Manuscriptorium using individual input interface, so called Connector, which is prepared according to the metadata properties.
- For Advanced digitization without the OAI-PMH interface: available metadata are processed using the individually prepared Connectors for partners with larger number of documents (where Connector creation is reasonable). The difference against OAI-PMH enabled partners is in the method of transferring the metadata to the input of Manuscriptorium.
- For Starting or smaller scale digitization projects: it is possible to use Manuscriptorium dedicated tools to create and transfer the metadata content. These tools are:
    - M-Tool: an application which enables to create the descriptive and structural metadata

  - M-Can: on-line application which enables upload of the metadata into the Manuscriptorium environment, check of correctness and subsequently transfer for import
- For Large-scale digitization projects without structural metadata: it is possible to create the metadata in an automated way (using some of the commonly available generation tools) and process the results as structural metadata within dedicated individual Connector.

## Users' Requirements for Community Building Features

As the resulting digital library service should be of high quality and user-centered, a survey was carried out investigating preferences of the digital library users in respect to static virtual collections, dynamic virtual collections and individual virtual documents which were the major functions planned for implementation in the user's personalization area. Individual collections allow any reader to create and maintain personal set of documents within reader's profile.
Virtual documents are documents created manually with parts of other digital library documents. For example, a reader can build a virtual document demonstrating the art of illumination for a selected period (e.g. showing all illuminations from one scriptorium in a virtual document in spite of the fact that they are from various originals owned by different institutions in different countries).
The survey was filled in by over 450 respondents from 12 European countries participating in the ENRICH project. Full analysis of the survey performed in frame of the ENRICH project is available in form of the ENRICH project deliverable D 4.1 [7]. The survey investigated necessity of the personalization features and it appeared that all of them were positively evaluated and therefore needed by the users. Third part of questions queried about the features such as possibility to attach a file or add notes to particular individual item (collection or document) were also positively evaluated. Functionality for sharing individual collections and documents with other users was welcome by most of respondents. The survey displayed that most of respondents prefer to choose which individual collections and documents will be available to other users and which will be hidden from them, that majority of respondents prefers to have the full configuration possibilities giving them full control over the shared content and that users would also appreciate possibility of giving a copy of personal collection or document to other users. The purpose of copying of an element is to create a new instance of the element and to enable selected user so he/she can start to use it as his/her own. It appears that for respondents the functionality to give a copy of an individual collection or document is more important than the functionality which allows to simply sharing it.

Another example of end-users interface extension is the Multilingual service module which above all enables to perform translation of selected records and also a pilot simple search query translation. The features enable the end-users to directly improve the translation results by enhancing the embedded historical documents specific translation dictionary.

## References


1. Knoll, Adolf – Mayer, Tomáš – Psohlavec, Stanislav – Vomlel, Jan: Digitization of Rare Library Materials. Storage of and Access to Data: The Solution for the Compound Document, Manuscripts and Old Printed Books [CD-ROM]. Praha: Národní knihovna České republiky, 1997.
2. Giesecke, Michael: Der Buchdruck in der frühen Neuzeit: Eine historische Fallstudie über die Durchsetzung neuer Informations- und Kommunikationstechnologien. Frankfurt am Main: Suhrkamp, 1998. 957 pp. ISBN 3-518-28957-8;
3. Uhlíř, Zdeněk: Teorie a metodologie elektronicko-digitálního zpracování rukopisů a hybridní knihovna. [The theory and methodology of electronic-digital processing of manuscripts and the hybrid library.] Praha: Národní knihovna České republiky, 2002. 324 pp. ISBN 80-7050-410-2.
4. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
5. Knoll, A., Digital Access to Old Manuscripts, Linguistica Computazionale, Digital Technology and Philological Disciplines, 277 – 286 (2004).
6. Mazurek, C., Stroiński, M., Węglarz, J., Werla, M., (2006). Metadata harvesting in regional digital libraries in PIONIER Network, Campus-Wide Information Systems, Vol. 23, No. 4, 241–253 (2006).
7. ENRICH project deliverable D4.1 – Definition of requirements for the creation of personalised virtual digital libraries, http://enrich.manuscriptorium.com/files/ENRICH_WP4_D_4_1_final.pdf
8. Roberto Caldelli, Cezary Mazurek, Paolo Mazzanti, Tomasz Parkoła, Marcin Werla: *Users requirements for personalised virtual digital libraries*, QQML2009 - Qualitative and Quantitative Methods in Libraries, International Conference, Chania Crete Greece, May 2009 http://www.isast.org/proceedingsQQML2009/PAPERS_PDF/Caldelli_et_al-Users_requirements_for_personalised_virtual_digital_libraries_PAPER-QQML2009.pdf

# Improving search in scanned documents: Looking for OCR mismatches

Alistair Willis[1], David Morse[1], Anton Dil[1], David King[1],
Dave Roberts[2], Chris Lyal[2].

[1] Department of Computing, The Open University, Walton Hall, Milton Keynes, UK
[2] The Natural History Museum, London, UK

Corresponding author: d.j.king@open.ac.uk

**Keywords:** Biodiversity, digital library, e-research, Needleman-Wunsch, OCR,

## Background

The ABLE (Automatic Biodiversity Literature Enhancement) project aims to enhance access to collections of scanned documents from the taxonomic literature. The older literature, dating from 15th century, can inform management practices in modern concerns, especially biodiversity loss, land-use patterns, sustainability and climate change. Therefore, unlike most other sciences, taxonomic research and usage require access to the full range and history of publications on the subject.

Biological taxonomy is the discipline that manages the names of living and fossil organisms, defining the relationships within and between them. It therefore provides the central infrastructure for information management in the biological sciences [1].

Publication through peer-reviewed journals is a relatively recent phenomenon. Until the 1930s, scientific observations appeared in a wide variety of publications, including learned societies (e.g. Proceedings of the Royal Society), Institutional annual reports (e.g. Verhandlungen des Naturwissenschaftlichen Vereins in Hamburg) and encyclopaedias (e.g. Bronn's Klassen und Ordnungen des Their-Reichs). Many of these publications are only held in a few libraries and are difficult to access. The difficulty of accessing taxonomic information is a severe impediment to research and delivery of the subject's benefits [2]. It has also been seen as a major impediment to implementing the Convention on Biological Diversity [3]. Taxonomic names change over time [4] and while this is both inevitable and desirable as knowledge advances, it makes information management more challenging. For example, the taxonomic hierarchies used by Catalogue of Life [5] and the National Center for Biotechnology Information [6] are different, so the collective groups that might be used in a search comprise different actual organisms.

## OCR and Terminological Variation

To liberate the information and data contained in the literature of the last 500 or so years, it is first necessary for these older publications to be digitised [7], for which industrial-scale scanning projects are essential. One such project is the Biodiversity Heritage Library (BHL) [8]. However, errors are introduced during the digitisation process because current OCR (Optical Character Recognition) technology is not perfect. The errors may mean that words are not recognised by standard search techniques, but at the current rate of scanning it is not practical to engage in manual validation and error checking of documents. To enable library users to search on the terms which are difficult for OCR systems to recognise, we therefore require mechanisms to reduce the impact of OCR errors.

This also applies to the task of automatic markup of taxonomic texts. Contemporary publications exploit the benefits of markup technologies for information sharing and information searching. Automatic markup of biodiversity texts will require accurate recognition of taxonomic names and then mark-up using extensions to existing XML schemas such as DjVu XML, SciXML [9] and NLM DTD (used by BHL). Ultimately the project will work towards full mark-up in the taXMLit schema [10]. We have already developed an XSL transformer for the reverse process, to extract source text from taXMLit documents.

OCR performs poorly on scanned pages, especially of older publications. These may have old typefaces and, to the modern eye, odd layout conventions [11]. Consequently, recognition accuracy is often worse than on modern publications. Errors introduced during digitisation give potential variations in recognised taxonomic names. For example, erroneous recognition of '*o*' in place of '*c*' might propose the taxon Pioa, not a known name, rather than *Pica* (European magpie). External data sources, e.g. Catalogue of Life and NameBank associate known latinised names with common names and synonyms, but being under active development, these are incomplete, and so cannot form the only basis for term recognition. In addition, mistaking '*o*' for '*a*' can change the genus *Homa* (a hemipteran insect) into *Homo* (mankind), so that non-appearance in an existing database cannot be used to identify errors. The BHL have found 35% of taxon names in scanned documents contain an error, with 50% of those errors being in one or two characters. Further, the genus name *Pieris* is a valid name for both a plant (*Ericaceae*) and a butterfly (including the cabbage white), so a single name can represent two quite separate concepts.

## Sequence Alignment to Identify OCR Errors

In order to start identifying some of the possible errors introduced by OCR, we are comparing the output of two different OCR packages. Modern OCR packages usually combine different feature-based as well as pattern matching classifiers and use internal voting to produce their final output. Differences between packages arise due to the dictionaries used and to the individual font recognition training. Each of these factors provides a challenge that the ABLE project will have to overcome. First, there

is no comprehensive dictionary of taxonomic names and second, in a distributed large-scale digitisation project such as the BHL, training the OCR packages with the multiplicity of fonts used in the source texts is impractical.

We have assumed that those terms which are difficult for an OCR package to recognise are those which are most likely to be interpreted differently by different packages. The outputs from the OCR packages are compared against a source document, drawn from a Biologia Centrali-Americana (BCA) volume which was used in the INOTAXA [12] project. This volume has been manually keyed in, and so is expected to contain (as far as possible) very few incorrect interpretations of the physical page. (INOTAXA found that manual rekeying of the journal content was more financially viable than automatic analysis of page scans.)

The text files we are comparing are both derived from a common PDF of the BCA volume. The first is taken from the Natural History Museum's work as part of the BHL, created with Adobe PDF maker and the associated OCR tool. The second was obtained from the Internet Archive [13] and was created using LuraTech PDF Compressor with ABBYY FineReader for the OCR.

To identify where the two OCR systems interpret strings differently, the two text files were split into words (using either whitespace or newlines as word separators), and compared using the standard Needleman-Wunsch algorithm [14] to align the texts. This algorithm performs a global alignment on two sequences by identifying the common terms between them, and inserting gaps or mismatches where no identical terms can be found. In our case, the mismatches identified by the algorithm are those terms which have been interpreted differently by the two OCR packages.

In practice, many of the misaligned terms are those that we would expect an OCR system to find difficult to recognise, and in fact, are often the taxonomic names that we would hope to recognise. Some examples are:

| Reference | ABBYY FineReader | PDF maker |
|---|---|---|
| Otiorhynchinæ | `Otiorhynchinse` | `Otiorhynchinae` |
| Epicærina | `Epicserina` | `Epicaerina` |
| Sciaphilina | `Sciaphilina` | `Sciaphiliua` |

showing that features such as ligatures cause problems for the OCR systems, but are not restricted to these (all the reference terms are italicised in the original document).

The comparison also highlighted other areas where the OCR systems return different results. The term '*RHYNCHOPHOBA.*' in the reference document was returned as '`BHYNCHOPHOKA.`' by PDF maker (illustrating some of the character misinterpretations), but as the pair of terms '`KHYNCHOPHOBA`' and '`.`' by ABBYY FineReader, illustrating both a spelling variant, and a different interpretation of the punctuation in the text. We do not currently analyse the punctuation in any way.

Our ongoing work is to identify how far the differences in the OCR outputs can be used to recognise the taxonomic names in the absence of a taxonomic dictionary to verify them, and whether it is possible to find systematic interpretations of the spelling variants that appear in these different outputs. This understanding can be used to clean up the OCR text should we be allowed to revise the published material, and if not then to enhance fuzzy searching of the text so that plausible variants are identified.

**Acknowledgements**

The work in this document is wholly funded by JISC, the UK's Joint Information Systems Committee.

**References**

1. Knapp, S., Lamas, G., Lughadha, E.N., Novarino, G.: Stability or stasis in the names of organisms: the evolving codes of nomenclature. Philosophical Transactions of the Royal Society. Series B, 359, 611–622 (2004)
2. Godfray, H.C.J.: Challenges for taxonomy. Nature. 417, 17–19 (2002)
3. SCBD: Guide to the Global Taxonomy Initiative. CBD Technical Series, 30, pp viii + 195 (2008).
4. D. M. Roberts.: Explaining taxonomy to kids. Society for General Microbiology Quarterly. 23(5) 7–8 (1996)
5. Catalogue of Life, http://www.catalogueoflife.org
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/Taxonomy
7. Curry, G.B., Connor, R.J.: Automated extraction of biodiversity data from taxonomic descriptions. Systematics Association Special Volume Series. 73, 63–81 (2007)
8. Biodiversity Heritage Library, http://www.biodiversitylibrary.org
9. Lewin, I.: Using hand-crafted rules and machine learning to infer SciXML document structure. In 6th UK e-science All Hands Meeting, National e-Science Centre, Edinburgh (2007).
10. Biodiversity Information Standards (TDWG) was known as the Taxonomic Database Working Group http,://wiki.tdwg.org/twiki/bin/view/Literature/WebHome
11. Lu, X., Kahle, B., Wang, J., Giles, L.: A metadata generation system for scanned scientific volumes. In 8th ACM/IEEE joint conference on Digital libraries pp. 167–176. IEEE Press, New York (2008)
12. INOTAXA ('INtegrated Open TAXonomic Access'), http://www.inotaxa.org/jsp/index.jsp
13. The Internet Archive, http://www.archive.org/index.php
14. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 48(3), 443–453 (1970)

Free University of Bozen-Bolzano
Faculty of Computer Science

www.unibz.it

ISBN 978-88-6046-030-1