Tiziana Catarci
Nicola Ferro
Antonella Poggi (Eds.)

Communications in Computer and Information Science 385

Bridging Between Cultural Heritage Institutions

9th Italian Research Conference, IRCDL 2013
Rome, Italy, January 31 – February 1, 2013
Revised Selected Papers

Springer

Communications in Computer and Information Science 385

Editorial Board

Tiziana Catarci Nicola Ferro
Antonella Poggi (Eds.)

Bridging Between Cultural Heritage Institutions

9th Italian Research Conference, IRCDL 2013
Rome, Italy, January 31 – February 1, 2013
Revised Selected Papers

Springer

Volume Editors

Tiziana Catarci
Sapienza Università di Roma
Rome, Italy
E-mail: catarci@dis.uniroma1.it

Nicola Ferro
Università di Padova
Padua, Italy
E-mail: ferro@dei.unipd.it

Antonella Poggi
Sapienza Università di Roma
Rome, Italy
E-mail: poggi@dis.uniroma1.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-54346-3 e-ISBN 978-3-642-54347-0
DOI 10.1007/978-3-642-54347-0
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: Applied for

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

Printed on acid-free paper

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

Preface

The Italian Research Conference on Digital Libraries (IRCDL) is an annual event for the Italian research community, both on the computer science and on the humanities side, interested in digital libraries, digital cultural heritage, and related topics. The IRCDL conferences were launched in 2005 by Maristella Agosti and Costantino Thanos and initially sponsored by DELOS, an EU FP6 Network of Excellence on digital libraries together with the Department of Information Engineering of the University of Padua. Over the years, IRCDL has become a self-sustainable event supported by the Italian Digital Library Research Community.

The focus of this ninth edition was on the multidisciplinary nature of research on digital libraries, which not only ranges from humanities to computer science but also crosses over areas in the same field ranging, for example, from archival to librarian sciences or from information systems to human – computer interaction. This is a continued challenge for the digital libraries field and there is the need to effectively bridge the gap existing between communities that share common objectives. The aim is therefore to provide the opportunity to explore new ideas, techniques, and tools, developed both in the humanities and computer science fields, and to exchange experiences from on-going projects.

This volume contains the revised accepted papers from among those presented at the 9th Italian Research Conference on Digital Libraries (IRCDL 2013), which was held at the Department of Computer Science of the Sapienza University of Rome, from January 31 to and February 1, 2013.

The recognized scope of IRCDL is to bring together the Italian research community interested in the diversified methods and techniques that allow the building and operation of digital libraries. A national Program Committee was set up composed of 15 members, with representatives of the most active Italian research groups on digital libraries.

The papers accepted for inclusion in this volume were submitted in an extended version with respect to the ones orally presented. Those papers underwent a new review process resulting in the contributions appearing in the volume. The covered topics are related to the different aspects that need to support information access and interoperability; among those there are:

- Formal and methodological foundations of digital libraries
- Digital library architectures and infrastructures
- System interoperability and data integration
- Ontologies and linked data for digital libraries
- Metadata creation, management, and curation
- User interfaces and visualization
- Information access, usability, and personalization
- Long-term preservation

- Collaborative environments
- Social networking and networked information
- Quality and evaluation of digital libraries
- Digital libraries for education and learning
- Digital libraries for the evaluation of research quality and scholarly impact (bibliometric indicators, citation analysis, ...)
- Exploitation of cultural heritage material

In addition to the presentations of the accepted papers, the program of IRCDL 2013 featured a keynote talk and a panel. The keynote talk by Paola Manoni of the Vatican Library was entitled "The digitization project of the Vatican Library within the complex relationships between sets of metadata" and focused on the metadata schemas involved in the digitization project of the Vatican Library: for long-term preservation strategies as applied to digital deposit collections, as well as for Web publication of images in the context of the digital library. The panel concerned the evaluation of cultural heritage information access systems, and was moderated by Tiziana Catarci, Sapienza University of Rome, while the panelists reflected the perspectives of different fields, ranging from computer science – Giuseppe Santucci, Sapienza University of Rome – to archival and librarian sciences – Mariella Guercio, Sapienza University of Rome, and Maurizio Messina, Marciana National Library – and digital humanities – Francesca Tomasi, University of Bologna.

IRCDL 2014, the tenth edition of IRCDL, will be held at the Department of Information Engineering of the University of Padua during January 30-31, 2014.

We would like to thank those institutions and individuals who made the conference and this volume possible. In particular, we would like to thank the Program Committee members and the additional reviewers, the Steering Committee members, the authors, the Department of Computer, Control, and Management Engineering Antonio Ruberti of the Sapienza University of Rome, the Department of Information Engineering of the University of Padua, CINECA, and the PROMISE FP7 Network of Excellence.

December 2013

Tiziana Catarci
Nicola Ferro
Antonella Poggi

Organization

IRCLD 2013 was organized by the Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome.

General Chair

Tiziana Catarci	Sapienza University of Rome

Program Chairs

Nicola Ferro	University of Padua
Antonella Poggi	Sapienza University of Rome

Local Organization Chair

Carola Aiello	Sapienza University of Rome

Program Committee

Michelangelo Ceci	University of Bari
Gianfranco Crupi	Sapienza University of Rome
Stefano Ferilli	University of Bari
Costantino Grana	University of Modena and Reggio Emilia
Maria Guercio	Sapienza University of Rome
Simone Marinai	University of Florence
Carlo Meghini	ISTI CNR, Pisa
Fausto Rabitti	ISTI CNR, Pisa
Marco Shaerf	Sapienza University of Rome
Gianmaria Silvello	University of Padua
Giovanni Solimine	Sapienza University of Rome
Carlo Tasso	University of Udine
Erilde Terenzoni	Soprintendenza Archivistica per il Veneto
Francesca Tomasi	University of Bologna
Paul G. Weston	University of Pavia

External Reviewers

Gianvito Pio	University of Bari

IRCDL Steering Committee

Maristella Agosti	University of Padua
Tiziana Catarci	Sapienza University of Rome
Alberto Del Bimbo	University of Florence
Floriana Esposito	University of Bari Aldo Moro
Carlo Tasso	University of Udine
Costantino Thanos	ISTI CNR, Pisa

Supporting Institutions

IRCDL 2013 benefited from the support of the following organizations:
Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome
Department of Information Engineering, University of Padua
CINECA
PROMISE FP7 Network of Excellence

Table of Contents

Keynote Address

The Digitization Project of the Vatican Library within the Complex Relationships between Sets of Metadata 1
Paola Manoni

Panel

Evaluating Cultural Heritage Information Access Systems (Panel) 7
Tiziana Catarci, Maria Guercio, Giuseppe Santucci, and Francesca Tomasi

Information Access

Contour-Based Progressive Identification of Known Shapes in Images . . . 17
Stefano Ferilli, Floriana Esposito, Domenico Grieco, and Marenglen Biba

EDB: Knowledge Technologies for Ancient Greek and Latin Epigraphy 29
Fabio Fumarola, Gianvito Pio, Antonio E. Felle, Donato Malerba, and Michelangelo Ceci

DL Architecture

Fostering Interaction with Cultural Heritage Material via Annotations: The FAST-CAT Way 41
Nicola Ferro, Gary Munnelly, Cormac Hampson, and Owen Conlan

EuropeanaLabs: An Infrastructure to Support the Development of Europeana 53
Nicola Aloia, Cesare Concordia, Carlo Meghini, and Luca Trupiano

A Digital Infrastructure for Trustworthiness: The Sapienza Digital Library Experience 59
Angela Di Iorio, Marco Schaerf, Maria Guercio, Silvia Ortolani, and Matteo Bertazzo

DL Projects

Historical Digital Archive and Geo-referenced Contents of the *Francigena Librari* Web Portal 70
Adriana Martinoli and Alfredo Esposito

The Heritage of the People's Europe Project: An Aggregative Data Infrastructure for Cultural Heritage 77
Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, Marko Mikulicic, and Franco Zoppi

Semantics and DLs

Towards a Methodology for Publishing Library Linked Data 81
Dydimus Zengenene, Vittore Casarosa, and Carlo Meghini

ConNeKTion: A Tool for Handling Conceptual Graphs Automatically Extracted from Text 93
Fabio Leuzzi, Stefano Ferilli, and Fulvio Rotella

Exploiting Wikipedia for Evaluating Semantic Relatedness Mechanisms 105
Felice Ferrara and Carlo Tasso

Semantic Lenses as Exploration Method for Scholarly Articles 118
Silvio Peroni, Francesca Tomasi, Fabio Vitali, and Jacopo Zingoni

Models and Evaluation for DLs

Digital Archives: Extending the 5S Model through NESTOR 130
Nicola Ferro and Gianmaria Silvello

Evaluation of Digital Humanities: An Interdisciplinary Approach 136
Anna Maria Tammaro

The Evaluation Approach of IPSA@CULTURA 147
Maristella Agosti, Marta Manfioletti, Nicola Orio, Chiara Ponchia, and Gianmaria Silvello

DL Applications

Optimizing Relevance Ranking to Enhance the User's Discovery Experience 153
Tamar Sadeh

Sapienza Libraries and Google Books Project 165
Adriana Magarotto, Maura Quaquarelli, and Mattia Vallania

Discussing DL Perspectives

Multimedia Digital Libraries Handling: The Organic MMIR Perspective 171
Roberto Raieli

Closing the Gap: Interdisciplinary Perspectives on Research and Education for Digital Libraries 187
Anna Maria Tammaro, Vittore Casarosa, and Donatella Castelli

Author Index 199

The Digitization Project of the Vatican Library within the Complex Relationships between Sets of Metadata

Paola Manoni

Biblioteca Apostolica Vaticana
manoni@vatlib.it

Abstract. The talk is focused on the metadata schemas involved in the digitization project of the Vatican Library: for long-term preservation strategies as applied to digital deposit collections, as well as for web-publication of images in the context of the digital library. The relationship management in the implementation process of sets of metadata (in their structural representation and semantic meaning of data elements) will be discussed with particular attention to the management implications and the resulting operational capabilities.

Keywords: Metadata, Digital Libraries, MARC21, TEI-P5, METS, PREMIS, FITS format.

1 Introduction

Metadata is the core of any information retrieval system and so its implications for any digital library are profound: the choice of a metadata scheme underpins any such library's ability to deliver objects in a meaningful way, and greatly affects its long-term ability to maintain and preserve its digital assets.

The overall goal of this presentation is to provide information about metadata infrastructures that affords interoperability among heterogeneous, autonomous digital library services implemented for the digitization project of the Vatican Library. These services include both search services and remotely usable information processing facilities.

Metadata required for a diverse set are surveyed and classified. Metadata architecture fits into our established infrastructure and promotes interoperability among existing and de-facto metadata standards. Several pieces of this architecture are implemented; others are under construction. The architecture metadata information offers facilities for search services, and local metadata repositories. In presenting and discussing the pieces of the architecture, we show how they address our motivating requirements. Together, these components provide, exchange, and describe metadata for information objects and metadata for information services.

2 Metadata Related to Objects

The digitization is performed for various projects operating in the library that relate to collections of manuscripts and incunabula.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 1–6, 2014.

Descriptive metadata are primarily used for resource discovery and format currently used include TEI-P5 format for manuscripts and MARC21 for incunabula.

The following summary describes metadata protocols involved in the interoperability between systems being implemented as **research tools** of the Vatican Library and the workflow planned for the forthcoming digital project.

3 Brief Description of the Existing Research Tools

3.1 Online Catalogues

The Vatican Library provides online catalogues to help researchers access and make better use of the collections.

There are separate online catalogues for each collection (Manuscripts, Archives, Printed books, Incunabula, Graphic prints and Drawings, Coins and Medals) and a General Catalogue able to perform interoperability between MARC21, EAD and TEI-P5.

- **Manuscript catalogue**: It includes complete or partial data taken from inventories, bibliographies, printed catalogues, card indexes. The encoding of the descriptive elements conforms to the **TEI specifications** and uses XML syntax;
- **Archival holdings catalogue**: It includes complete or partial data taken from inventories. The encoding of the descriptive elements conforms to the **EAD specifications** and uses XML syntax;
- **General Printed books catalogue**: It includes the description of the entire collection of printed volumes (monographs and periodicals) from the XVIth century to the new acquisitions. Cataloguing is carried out in MARC 21 format;
- **Incunabula catalogue**: It includes bibliographic records related to the VISTC (*Vatican Incunabula Short Title Catalogue)* and the *BAVIC*[1] of the entire collection where links to persistent URIs related to the digitized volumes are provided. Cataloguing is carried out in MARC 21 format;
- **Graphic prints and Drawings catalogue**: It includes the descriptions of the prints, maps, drawings, photographs and plates which are kept in the various collections of the Library. Cataloguing is carried out in MARC 21 format;
- **Coins and Medals catalogue**: It includes descriptions of the coins and medals kept in the Library. A running project aims to make digital scans of the graphic prints, coins and medals, insert hypertextual links to them into the related bibliographical descriptions, and enter the data into each web-based catalogue of the Library Cataloguing is carried out in MARC 21 format.

3.2 Systems in Use

The online catalogues use two main systems for the management of different metadata: TEI-P5 and EAD in XML syntax for manuscripts and archival units (in two

[1] The project BAVIC (Bibliothecae Apostolicae Vaticanae Incunabulorum Catalogus) is the analytical cataloging of 8,600 incunabula.

separate collections of data but in the same application named InForMA, entirely developed at the Vatican Library using open-source Java/XML technology for data archive, authority indexes and search engine); and MARC21 for the other collections (with different specifications for each type of document).

- **InForMA** — The implementation provides: full native XML database support, storing XML content as is and providing true XML retrieval capabilities based on the XPath and XQuery standards; handling of structured and unstructured multimedia rich content (it can store and retrieve multiple different file types); integration with Microsoft Office and other WebDAV-enabled product suites. These features are supported by the embedded **Tamino** technology provided by the German Software AG company.

  The implementation of the TEI schema for manuscripts also provides a kind of information that includes data element as atomic units applied for internal usage in which the persistent URI related to the web-presentation for each digitized manuscript is given.

- **V-Smart / Iguana** – It allows scholars to query either the integrated general catalogue, where they can have a quick and thorough search result related to any bibliographic resource, or each catalogue where bibliographic descriptions are stored in their native format. The technological infrastructure is based on OAI-PMH protocol and script elements for exporting data (from InForMA) and importing XML (to V-Smart). From each bibliographic record, V-Link (openURL resolver) is able to search and access a variety of information resources and retrieve truly relevant search results.

  The system is able to support the persistent URIs related to the web-presentation of the digitized incunabula.

4 Structural Metadata

Information necessary to record the internal structure of an item so that it can be rendered to the user in a sensible form. This type of metadata is necessary as an item may often be comprised of multiple of the images of individual pages that make up a digitized book.

For the digital project the Vatican Library has adopted METS standard automatically created by the use of the DWORK[2] implementation, provided by the UniversityLibrary of Heidelberg to support the process of digitization and the web presentation of the digital objects.

[2] The University Library of Heidelberg in-house development. It supports the process flow of digitization and the web presentation of the digitized works.

The software as a web application thereby supports all single steps of the workflow from the creation of metadata, scan processing, creation of the web presentation to the storage of scans and metadata.

5 Administrative Metadata

They provide information about the management of the digital collection and facilitates long-term management and processing of it. They include:

- quality control, rights management, access control and use requirements;
- technical data on creation (such as scanner type and model, resolution, bit depth, color space, file format, compression, light source, owner, copyright date, copying and distribution limitations, license information, preservation activities (refreshing cycles, migration, etc.);
- preservation, action information.

While the first point in list pertains to the METS file generated for each digitized unit/volume, preservation and action information are managed within the PREMIS framework.

The use of the PREMIS is closely related to the concept of long-term preservation that the Library has adopted.

The photo workflow includes the acquisition of images in tiff/ raw formats. These files are used for the generation of the web-presentation procedure in the DWORK (METS and set of JPEG files). Then images are converted into FITS format and stored into the certified WORM data storage device for long-term preservation[3].

In the context of a discussion about metadata and how they interact, the consideration on the choice of storage formats has no place but, in the use of PREMIS in the Vatican Library, it is important to highlight the link between the information concerning the technical metadata format of the converted images and the use of the extension container 'object Characteristics Extension' that gives a place to record technical metadata defined by the FITS dictionary.

In fact a FITS file is made of 2880-byte records called 'FIT*S* blocks' divided between a header and a data area.

Each FITS file consists of one or more headers containing ASCII card images (80 character fixed-length strings) that carry keyword/value pairs, interleaved between data blocks. The keyword/value pairs provide information such as size, origin, binary data format, free-form comments, history of the data.

In the occurrence of the long-term archiving of the digital image, an automatic procedure extracts the embedded data in the header and defines the above mentioned

[3] Manuscripts and antique books of the Vatican Library's collections are being digitized to preserve them for future generations adopting a file format developed in the context of space missions and storing satellite images of the sky during the end of 70s of the last century: the Flexible Image Transport System (FITS), an open format, fully documented, without royalties or copyright, based on a series of specification publicly available and managed by a non-profit scientific authority.

The conversion TIFF/FITS takes into account all the characteristics of technical metadata (for example in TIFF format) in order to identify matches in the keywords of the header portion of the FITS file.

extension of the PREMIS. The structure of PREMIS is then completed with the main features related to the digital object and semantic units are defined to record the *environment* of each object.

6 Interoperability

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

The interoperability and exchange of metadata is further facilitated by metadata crosswalks. A crosswalk is a mapping of the elements, semantics, and syntax from one metadata scheme to those of another.

The workflow management of metadata, involved in each phase, from the acquisition of images to the archiving of the digital object, includes several steps and specific softwares:

1. Filename of TIFF / RAW file: An application able to assign a highly structured filename has been implemented[4].
2. The file name is automatically interpreted by the above mentioned DWORK so that the logical and physical sequences for each file group are added in the METS file related to each unit/volume.
3. Structural metadata that contain a table of contents with links to key structural elements such as title pages, table of contents, chapters, parts, sections and subsections (depending on the item) are added and automatically converted in the METS file.
4. Essential descriptive metadata are given in the MODS section of the METS file. Crosswalks between MODS and TEI-P5 / MARC21 has been established.
5. METS files related to manuscripts are exported from the DWORK to be processed by a specific console application in InForMA where the URI of the web presentation in the element 'FLocat LOCTYPE'is treated as an *ad hoc* element TEI-P5 document through specific Xquery functions.
6. METS related to incunabula are exported from DWORK in order to extract the 'FLocat LOCTYPE' for the creation of the link in the above mentioned OPACs, for each MARC21 bibliographic record. The information about the description of the digital object is performed in a qualified DC record in the OPAC.
7. TIFF format is converted in FITS format and a crosswalk between embedded technical metadata of TIFF and keywords in FITS header has been implemented.
8. PREMIS entities described in XML files are added at the time of archiving digital objects in the WORM storage device.

[4] Programming was carried out by specialists of the company Metis Systems s.r.l.

9. End-users can query the Web OPACs and get information about digitized manuscripts and incunables or browse the list of shelfmarks for each digital collections available in the website.
10. Reproductions of digital object may be requested. Queries are performed in the PREMIS database and a conversion from FITS file to an exchange format is available for private study or professional use.

References

1. Manoni, P.: Metadata framework and application profiles in the global structure of catalogs and digitization projects of the Vatican Library. Global Interoperability and Linked Data in Libraries: Special Issue. JLIS 4(1) (2013), http://leo.cilea.it/index.php/jlis/issue/view/536
2. Manoni, P.: L'interazione tra banche dati e analisi dei modelli descrittivi nella Biblioteca Apostolica Vaticana. In: Il Libro Antico Tra Catalogo Storico e Catalogazione Elettronica. Convegno internazionale, vol. 127. Accademia nazionale dei Lincei, Roma (2012)
3. Manoni, P.: The Vatican Library Web-based application for managing manuscript metadata stored in native XML databases. In: Current Research in Information Sciences and Technology, Proceedings of the Ist International Conference on Multidisciplinary Information Sciences and Technologies, Badajoz, Merida, vol. 2, pp. 570–575 (2006)

Evaluating Cultural Heritage Information Access Systems
(Panel)

Tiziana Catarci[1], Maria Guercio[1,2],
Giuseppe Santucci[1], and Francesca Tomasi[3]

[1] Sapienza Università di Roma, Italy
{catarci,santucci}@dis.uniroma1.it, maria.guercio@uniroma1.it
[2] Digilab, Roma, Italy
[3] University of Bologna, Italy
francesca.tomasi@unibo.it

1 Introduction

IRCDL[1] is a yearly deadline for Italian researchers on Digital Libraries related topics. This year the focus of IRCDL 2013 was on emphasizing the multidisciplinary nature of the research on digital libraries which not only goes from humanities to computer science but also crosses among areas in the same field ranging, for example, from archival to librarian sciences or from information systems to human-computer interaction.

The panel on "Evaluating Cultural Heritage Information Access Systems" was also characterized by this interdisciplinary flavour. Indeed, the panelists reflected the perspective of different fields, ranging from computer science (Prof. Giuseppe Santucci), to archival and librarian sciences (Prof. Mariella Guercio and Dott. Maurizio Messina) and digital humanities (Dott.ssa Francesca Tomasi).

It is evident that in these last years there is a growing and persistent demand for more and more digital content in many different areas and for diverse purposes, with a particular emphasis in the cultural heritage sector. Typically, raw digital content is assembled in digital collections, but only the curation and enrichment of the raw material make it usefully available for working with it and exploiting it, and this subsequent embellishment phase is carried out in digital libraries. Digital libraries basically consist of large digital collections plus a set of tools that make content alive, that help the users to find it, make sense out of it, annotate it, comment it, share it in a community, collaborate on it, and so on. Thus, the evaluation of digital libraries and information access systems has to consider not only the quality of the digital collections and the preservation approach, but also the effectiveness and efficacy of the user-oriented tools they provide, in an overall user experience, having the ultimate goal of creating new knowledge.

Following the above idea, the panelists were presented with four specific topics:

– Cultural Heritage Information Access System vs Digital Library: similarities and differences

[1] http://ims.dei.unipd.it/websites/ircdl/home.html

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 7–16, 2014.

- Accessing Raw Material vs interpretation and curated presentation of such material
- Dimensions of the User Experience
- Evaluation coordinates: quality and completeness of the information, tools to make sense out of it, usability, accessibility, open access,

And asked to give their view with respect to one or more of them. Summaries of their presentations are presented in the following sections.

Finally, the panel - and IRCDL 2013 in general - would not have been possible without the generosity of our sponsors, namely Sapienza Universit di Roma, PROMISE (Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation - Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191) and the CINECA Consortium.

2 A Contribution from the Archival Domain

The presentation has focussed the attention to only one specific aspect of the complex issues under discussion in the panel: the capacity of representing the archival cultural heritage when the representation has to be included as part of a large and multidisciplinary digital library. In particular the presentation has analysed the challenges which always emerge when a cross-domain heritage is involved and variegated and rich contexts and provenance have to be described and, even more, rendered in large web environments. This aspect have been considered from the perspective of the archival heritage and its main (and not avoidable) complexity but, as mentioned, the presentation has not underestimated the general need for cultural heritage collections (of any kind) to be represented and made available within their original structure and relations.

From this point of view, the archival attention for the provenance and the contextual information can provide a fruitful methodology for future qualified solutions, for a better structured information representation and for a higher degree of intelligibility of the digital resources made available on the web. As discussed in the panel, the limits of the present evolution is partially due to the lack of participation of digital curators other than librarians in the processes related to the communication processes of digital heritage: the archivists but also the professionals active in the cultural heritage sector have not realized the strategic relevance of this cooperation and have not sufficiently contributed to the definition of a more comprehensive integrated approach to the digitization process and to the interoperability of descriptive metadata. The consequences of this absence have been: too granular and flat solutions, no capacity of collecting relations and supporting contexts and provenance information, a very limited number of innovative proposals in the field. Among others, the presentation has discussed:

- the *key questions* still unsolved for defining the quality of a digital library according to the DELOS requirements like a not yet well defined assessment framework, the lack of attention for an efficient capacity of retrieving and

managing digital library contents when peculiar representations and information structures have to be supported,
- the limits of the available solutions to support the *complex translation into the web of archival finding aids systems* and resources with reference to their uniqueness and to the complexity of their reciprocal relations (strategic for making them understandable), but also in consideration of the massive volumes (approximately 8000 km in Italy of unique resources) which limit the role of the digitization process in this area (the finding aids system will necessarily focus, at least for the next decade, on analogue records to be accessed in the traditional reference services and only few resources will have the "privilege" to be digitized), the risk of fragmentation and arbitrary criteria for selection: the thematic approach generally supported by the digital library environment can be useful for attracting users but not for providing qualified and inclusive services to the scientific communities and to the citizens needs,
- the *ambiguity of the concepts involved* which has prevented the archival sector by using even the terms involved in this new environments: archival websites or archival information systems are the preferred terms and the preferred scenarios for developing services and functionality, while the term digital library is at the moment rarely used for proposing online publication of digital archival resources in dedicated environments.

For providing concrete and technical elements for a more detailed analysis, the presentation had considered best research projects and, if possible, good practices already available at national level to sustain future cooperation in this field (just as examples of promising services for future development and a more comprehensive analysis). They have included the following cases:

- SIAR (Sistema informativo archivistico regionale del Veneto), created with the aim of integrating distributed archives by using both standards, methodology and tools developed for digital libraries and for specific domain, but also for exchanging or sharing metadata, even if, at the moment, the integration for research services is not explicit and the catalogues are maintained separate,
- Biblioteca digitale della Lombardia, whose aim is the digitization and online publication of cross-domain cultural resources able to testify the regional cultural heritage and whose basic requirements are: the acquisition and online availability (through a regional web portal) of significant resources of regional culture, the integration of resources preserved and created in different environment, including archival documents, the capacity of ensuring the long-term preservation of digital resources, in compliance with the main standards and in cooperation with the main national and international projects (Europeana, CulturaItalia and Sistema archivis-tico nazionale),
- Sapienza Digital Library, whose ambitious goals are: 1) the aggregation and accessibility in a digital form of cross-domain information contents created by Sapienza University research communities or made available by corporate bodies or individuals in relation to the academic environment; 2) the

harmonization of the descriptive practices for new resources (not easily ascribable to a specific domain) by adopting with some degree of "creativity and imagination" national standards and recommendations with reference to the use of controlled vocabularies, ke PICO (Portale della Cultura Italiana) 4.3 MibAc and *Nuovo soggettario di Firenze* for subjects, TGN (Thesaurus of geographic names) Getty and Geonames, but also VIAF Virtual International Authority file and other internationally recognized vocabularies based on specific disciplines.

The critical state of art of the Italian projects has been also briefly examined and some conclusions made, as here summarized:

- provenance and context are not always identified as crucial components;
- archival standards are recognized for their general value, but not yet completely implemented outside the archival information systems,
- compliance with European standards is generally stated and partially ensured but mainly as a static and flat model for representation,
- the main difficulties concern: the differentiation of digital resources (many projects are limited to the identification of single resources and have developed simple research interfaces), the low level of integration and cooperation among institutions both at regional and at national level and of course and the lack of financial resources
- the integration (and not convergence) among heterogeneous cultural information access systems is here the key term, but a balance is required between specificity, details and general perspective: the functionality for retrieval must be easy to use but not trivial and new forms for intermediation are required (particularly when digital resources are complex and articulated).

3 Evaluating Cultural Heritage Information Access Systems through Visual Analytics

The evaluation of an Information Access System is a non trivial task and several research activities focus on supporting it (see, e.g., *Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE)*). Different systematic methodologies and techniques are used in such a context and all of them share a common outcome: they produce a lot of complex data, whose interpretation is a challenging activity by itself. Lessons learned point out some key areas that can improve the overall process:

- Building a community focused on common objectives: that allows for better shaping methodologies and techniques, sharing ideas, experiences, and test cases. Moreover that allows for a sound and productive comparison of evaluation results;
- Using a robust infrastructure: computer support is mandatory for managing and providing access to the scientific data produced during evaluation activities, with the final goal of supporting the organization and the execution of evaluation activities, increasing the automation in the evaluation process;

– Visualizations: suitable visual analysis can foster and improve the usage and the interpretation of the managed evaluation data. Visualizations must be integrated in the system and an overall methodology like *Visual Analytics (VA)* is needed.

VA [1] is an emerging multi-disciplinary area that takes into account both ad-hoc and classical *Data Mining (DM)* algorithms and *Information Visualization IV (IV)* techniques, combining the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where human beings and machines cooperate using their respective distinct capabilities for the most effective results. Decisions on which direction analysis should take in order to accomplish a certain task are left to the user. Although IV techniques have been extensively explored [2], combining them with automated data analysis for specific application domains is still a challenging activity [3].

In order to apply VA techniques to *Cultural Heritage Information Access Systems (CHIAS)* evaluation, improving the visualizations, analysis, and interpretation of experimental data, it is mandatory to understand the structure of the data that are actually used for evaluation purposes. A full discussion of this aspect is out of the scope of this panel. Here we report some results coming from the PROMISE NoE project, whose experimental datasets rely on the idea of evaluation campaigns and have a quite broad validity.

An evaluation campaign provides large test collections (like multimedia, multilingual, text, images), specifying a set of topics and a relevance assessment.

Participants evaluates their searching algorithm(s) (i.e., experiments) against a specific collection, producing a (ranked) result set on which different metrics are computed and stored. These data can be represented by the TME (Topics-Metrics-Experiment) cube shown on Figure 1.

Starting from this cube, it is possible to aggregate or manipulate data in different ways, according to different evaluation needs that, roughly speaking,

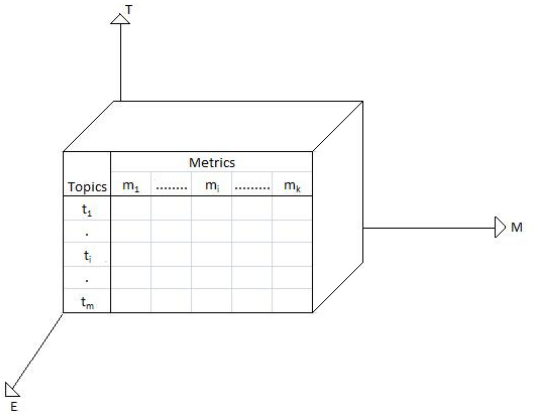


Fig. 1. The TME Data cube

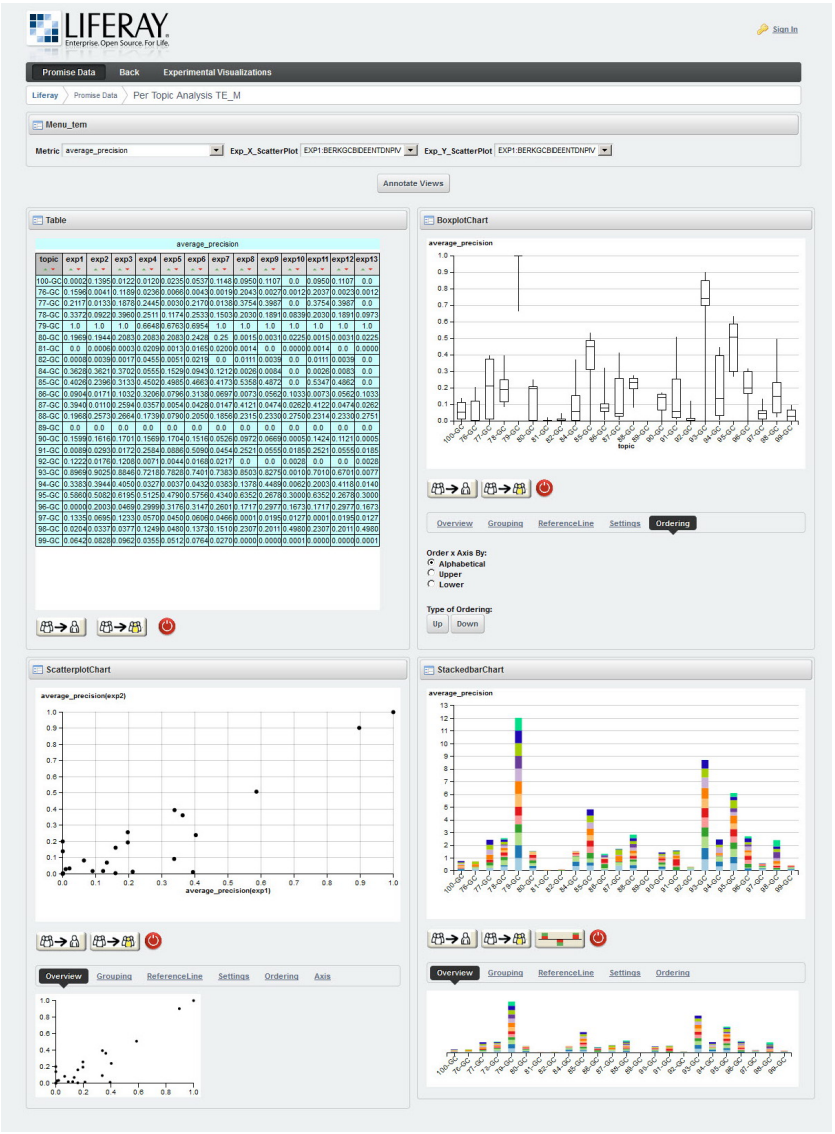


Fig. 2. Per topic analysis

correspond to a) evaluating a single search engine, either topic by topic or as a whole or b) to compare two or more search engines.

According to these two evaluation analysis patterns a VA prototype has been developed within PROMISE, implementing two visual analysis patterns, namely *Per topic analysis* and *Per experiment analysis*. In the following we describe such a prototype to provide a practical example of how a suitable visual analysis can foster and improve the usage and the interpretation of the managed evaluation data.

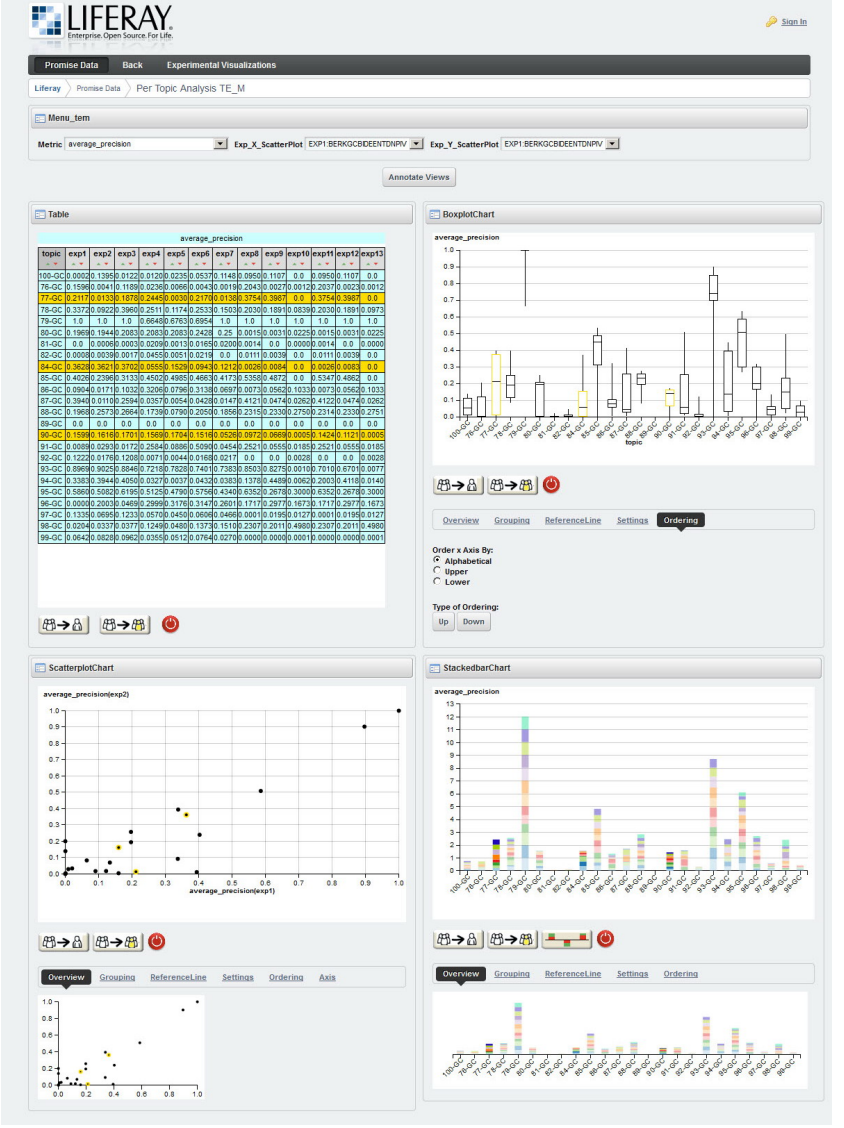


Fig. 3. Per topic analysis: a highlight operation

Per topic analysis Per topic analysis allows for comparing a set of search engines on each topic with respect to a chosen metric. Therefore the first step for an evaluator is to choose a metric m and, because the analysis implies a comparison on each topic, we represent topics on the x-axis in each available visualization. We foresee four coordinated visualizations: a table, a boxplot chart, a bi-dimensional scatter plot, and a stacked bar chart.

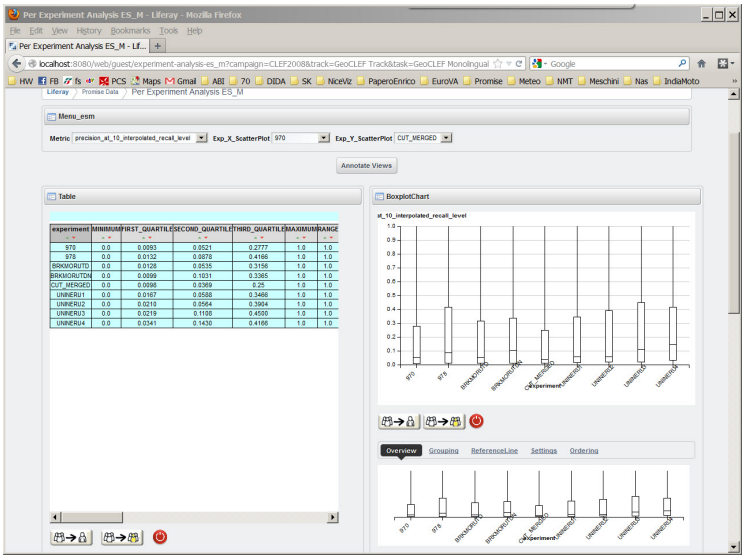


Fig. 4. Per experiment analysis: table and box plot

- The table represents a projection of the the TME(m) cube on the chosen metric, allowing for looking up details (see Figure 2, upper left);
- The box plot chart (see Figure 2, upper right) is used to evaluate the trend of topics among experiments with respect to the chosen metric m. It contains a box plot for each topic (x-axis) and the chosen metric (y-axis). Each box plot is built calculating statistical indicators on the set of data represented by a single TE(m) row;
- The bi-dimensional scatter plot (see Figure 2, lower left) allows for comparing topic behavior with respect to experiments. Each topic is represented by a point, according to the values it shows on the two experiments.
- The stacked bar chart (see Figure 2, lower right) has the same purpose as the box plot chart: to evaluate the trend of a topic among experiments with respect to a chosen metric. In such a visualization each topic is associated with all the values the explored metrics m exhibits in all experiments and the height of the bar represents the sum of all these values.

The evaluator can change the metric under analysis and restrict his or her focus on data subsets. As an example, Figure 3 shows three topics highlighted in all the four visualizations.

Per experiment analysis Per experiment analysis allows for analyzing a search engine as a whole and/or comparing the performances of a set of search engines with respect to a chosen descriptive statistics. As an example, on Figure 4, left side, the table chart represents a search engine in each row, showing the descriptive statistics of the metric average_precision (min, max, median, etc.). The box plot chart on Figure 4, right side, shows the percentile values of the observed metric for each experiment represented through boxplots.

Summarizing, we have introduced VA, a new, challenging metodology for analyzing large and complex dataset showing how apply it in the context of the evaluation of CHIAS, dealing with data structure and visualization requirements. We have used the European NoE PROMISE and the *Cross-Language Evaluation Forum (CLEF)* conference series as bed tests, but the ideas and the results presented in this panel have a quite broader scope.

4 Modeling Heterogeneous Humanistic Data in a User-Oriented Perspective

The primary step in the evaluation of a cultural heritage information access system from a digital humanist researchers point of view, consists in state clear what Digital Humanities (DH) is. DH is the name given to the alliance between associations involved in the study of relations between humanities and computer science (ADHO (Alliance of Digital Humanities Organizations); DH is a word referring to the main conference[2] in the field; DH is a open-access peer-reviewed electronic journal (Digital Humanities Quarterly[3]); DH is the title of publications, centers, blogs, projects and tools (see the "guide to digital humanities & arts"[4]). DH is a label that only recently started to identify a field historically known as "humanities computing" or "humanities computer science"; that is "computing" was replaced with digital. This terminological shift [4] corresponds to a new method of thinking about cultural objects as resulting from media integration. We assisted to the migration from a text-based computability process (i.e. markup languages, vocabularies and schemas like TEI, text analysis, text mining, text processing) to a process of representation and description of each informational resource in a social dimension (images, audio and video and their integration in computational systems for information retrieval/extraction/mining).

The purpose of DH is conceptualization, modelization, formalization of humanistic data/content (that is the domain); the goal of DH is to define methods and develop strategies of domain representation in order to computability. In this perspective the Cultural Heritage (CH) is a multimedia humanistic domain that requires to deal with formats and data types, to manage metadata and controlled vocabularies, to define procedures regarding these data selection and dissemination. The information access system, instead, is connected to Digital Library (DL) and regards the choice of the infrastructure, the definition of the services offered, the digital objects (in this case deriving from the CH domain) and the final users [5].

In this scenario, DH is procedure of modeling heterogeneous data, that is of managing information organization methods [6], in order to produce complex digital objects with the aim to acquire knowledge. This means that DH works by abstraction, enucleating the hermeneutical pertinent elements of a collection, for what concerns the computational objectives; this abstraction requires then to

[2] http://adho.org/conference
[3] http://www.digitalhumanities.org/dhq/
[4] http://www.arts-humanities.net

reflect on the data model and finally on the information architecture, i.e. user access/interface and user needs.

In this context the user has a crucial role. A good data model potentially allows a good user access. Reading, browsing, tagging, query are examples of services offered by the DL; they depend both on the modelization at information system level and, consequently, the DL access criteria. The browsing method is one of the main evaluation methods of a CH information access system. In particular, browsing by relationships is an essential issue. It depends on both the data model and the user access. The data model, that is the ontological representation of the domain, is fundamental, because semantic relationships constitute the most exhaustive exploration method. Different levels of relationships, semantically declared, could be established: lexical networks (hyponyms; hyperonyms; synonyms; meronyms; related terms), structural connections (intertext, paratext, hypertext, architext, metatext, following the Genette classification [7]), concepts and topics (overlapping meanings, discipline and the object of study, persons and occupations, products and institutions). A good CH information system has to provide methods for acquiring knowledge through browsing concepts explicitly related at a certain level. The user access finds a method of exploration of the modeled collection by a "faceted navigation" [8]. Facets represent classes and the data filtering is expression of the predicates. The data model could find a classification output through facets, that control heterogeneous data expressed by different media.

One last consideration. Both the data model and the facets are expression of a specific point of view. The interpretation of the collection not only determines the user satisfaction but also fulfills the user needs. User log is a gold mine useful to understand effectiveness and efficiency of an information system both at data model level as well as at navigation behavior level.

References

1. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 154–175. Springer, Heidelberg (2008)
2. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, pp. 336–343 (1996)
3. Keim, D.A., Kohlhammer, J., Santucci, G., Mansmann, F., Wanner, F., Schaefer, M.: Visual analytics challenges. In: Proceedings of the eChallenges 2009 (2009)
4. Svensson, P.: Humanities computing as digital humanities. Digital Humanities Quarterly (2009)
5. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.P., Kovács, L., Landoni, M., Micsik, A., et al.: Evaluation of digital libraries. International Journal on Digital Libraries 8, 21–38 (2007)
6. Taylor, A.G., Joudrey, D.N.: The organization of information. Libraries Unlimited, Westport (2004)
7. Genette, G.: Palimpsestes. Seuil, Paris (1982)
8. Broughton, V.: The need for a faceted classification as the basis of all methods of information retrieval. In: Aslib Proceedings, vol. 58, pp. 49–72. Emerald Group Publishing Limited (2006)

Contour-Based Progressive Identification of Known Shapes in Images

Stefano Ferilli[1], Floriana Esposito[1], Domenico Grieco[1], and Marenglen Biba[2]

[1] Dipartimento di Informatica, LACAM Laboratory
Università degli Studi di Bari "Aldo Moro"
{stefano.ferilli,floriana.esposito}@uniba.it
[2] Computer Science Department
University of New York, Tirana (Albania)
Marenglenbiba@unyt.edu.al

Abstract. Information Retrieval in digital libraries is at the same time a hard task and a crucial issue. While the primary type of information available in digital documents is usually text, images play a very important role because they pictorially describe concepts that are dealt with in the document. Unfortunately, the semantic gap separating such a visual content from the underlying meaning is very wide, and additionally image processing techniques are usually very demanding in computational resources. Hence, only recently the area of Content-Based Image Retrieval has gained more attention. In this paper we describe a new technique to identify known objects in a picture. It is based on shape contours, and works by progressive approximations to save computational resources and to improve preliminary shape extraction. Small (controlled) and more extensive experiments are illustrated, yielding interesting results.

Keywords: Shape Recognition, Information Retrieval, Document Processing, Digital Libraries.

1 Introduction

Graphical components are a precious source of information to understand, index and retrieve documents in a digital library based on their content. Indeed, the power of modern technology allowed to efficiently and effectively store documents that are not just made up of text, but include (often a significant amount of) pictorial content whose relevance to the document cannot be ignored. Accordingly, while much effort was devoted in the past decades to information extraction from textual components, more recently significant attention has been paid towards images, as well.

Computer Vision deals with the analysis of digital images by computers, in order to discover and understand what is represented therein, and where. While vectorial images explicitly represent shapes and other geometrical elements, raster images pose the additional problem that no high-level information is available therein, and each pixel is syntactically (although, clearly, not semantically) unrelated from all the others. An important sub-field of Computer

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 17–28, 2014.

Vision is Object Recognition (*OR*) [5], that has many applications in automation processes. Recognizing an object means being able to distinguish it from a set of other objects. OR techniques usually classify objects based on distinguishing characteristics of the class they belong to, extracted from the image through a sequence of pre-processing steps. This requires to preliminarily analyze a set of objects of a known class to acquire the most relevant information to be subsequently exploited. However, understanding an image does not mean just being able to retrieve other images in a database that are pictorially similar to it; it also involves recognizing what that image is about, including (or starting from) the objects it contains. Content-Based Image Retrieval (CBIR) [2] focuses on image content, rather than on their overall features.

This work concerns Object Recognition aimed at understanding raster images by looking for known shapes in them. It proposes a combination of existing and novel image processing techniques, as a preliminary step to describe images using higher-level, human-understandable concepts and relationships among them. Such descriptions might be exploited as metadata to be added to the documents where the image appears, or be input to standard text processing techniques in order to index the documents based also on their pictorial content, or be fed to relational Machine Learning systems to infer models of image classes when the depicted information is too complex for standard propositional and statistical techniques. Although a full semantic understanding of the image meaning is still to come, this might nevertheless bring many advantages, among which the possibility of retrieving and relating documents in a digital library according to the images they contain, and providing explicitly otherwise implicit and latent information. This will also allow to perform queries using visual information such as images in addition to standard textual search techniques (e.g., by providing a sample image expressing the concept to be searched for).

The focus of this paper is on the overall technique and on its performance, rather than on the details of its single steps. In particular, here we present novel experimental results on the technique originally presented in [4]. After recalling some background notions and related work in next Section, the proposed technique will be described and evaluated in Sections 3 and 4, respectively. Lastly, Section 5 will conclude the paper and outline future work issues.

2 Background and Related Work

Although the techniques and algorithms to perform automatic Object Recognition are very different, depending on the operating environment, they all rely on a common background made up of image processing techniques, and follow a general workflow consisting of three steps [7]:

1. Image Processing: transforms the source image in another image more suitable for running subsequent steps and reaching the objectives;
2. Feature Detection: applies methods aimed at extracting characterizing elements of an image that are more significant than single pixels;

3. Recognition: exploits the features extracted in previous steps to first define classes of objects and then retrieve objects belonging to those classes.

Concerning step 1, a raster digital image consists of a set of primitive numeric items (pixels) that in isolation provide little significant information, just like a single element of a puzzle does not allow to understand the meaning of the whole picture. Several pixels, taken together, may make up more significant items such as lines, contours, blobs, textures. To be able to extract such a kind of information, often the image must be properly pre-processed using particular *filters*, i.e. functions operating on pixels that enhance some important details and/or dim other, less significant ones, such as the noise introduced by the acquisition means or by the representation format (if lossy).

Step 2 identifies and extracts significant information from the pre-processed image resulting from (a combination of) the aforementioned techniques. The information obtained in this way allows for a higher-level interpretation of the image. Depending on the kind of features to be extracted, several techniques are available, and often specific features are exploited for particular objectives.

As regards step 3, each element identified in the image can be compared to previously stored models in order to check possible correspondences. This is done by different algorithms, considering different kinds of information. Limitations in applying Computer Vision systems come from the difficulty in extracting information from images. For an Object Recognition system to be effective and flexible, several properties are desirable. Here, we focus on the following ones, deemed as very important [1]:

- Scale invariance.
- Translation invariance (the position of the object to be recognized cannot be assumed to be fixed in the acquired image).
- Robustness to change in intrinsic variables of the image (even in controlled environments, small changes in color, luminance or contrast can take place).
- Rotation invariance. Unfortunately, rotating a 3D object usually results in completely different shapes depending on the perspective; nevertheless, making the system robust at least to 2D rotation already ensures a noteworthy degree of reliability.
- Efficiency (usually in contrast to effectiveness).

While several proposals are present in the literature to face these problems (e.g., [6]), here we refer to the technique described in [4]. In that work, the identification of potential objects in the image, their representation and storage in suitable data structures and a corresponding matching algorithm that allows to detect known objects in new images were first introduced and described.

3 Object Recognition Technique

The object recognition technique proposed in [4] aims at identifying regions of an image that correspond to objects, and at recognizing the class of these objects

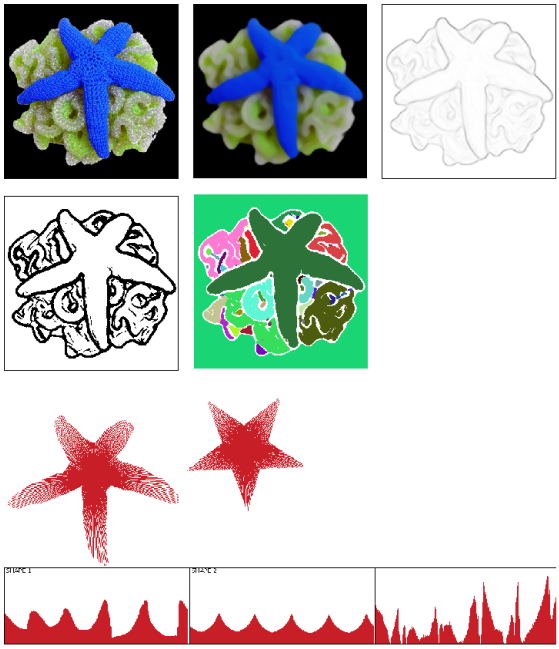


Fig. 1. Processing steps on a sample image

in a simple and effective way. While regions are determined on the grounds of color homogeneity, class recognition exploits the region contours. A graphical summarization of the various steps, applied by this technique to the original image in the top-left, is provided in Figure 1, while next subsections provide a description and discussion of these steps.

3.1 Pre-processing

We would like to blur the image within objects, so that they can be considered as single blobs by the segmentation step, but without blurring (and possibly even sharpening) their contours also, otherwise the resulting shape would be meaningless. Then, we need to extract the single blobs by preserving their contour peculiarities. Although any blurring and edge detection technique available in the literature can be used in this step, we purposely developed two filters that better reach these objectives, whose results are shown in the top row of Figure 1.

The image segmentation step, shown in the second row from the top in Figure 1, aims at identifying the candidate objects represented in the image. In this case we used standard binarization and region growing. As to the former, we used tresholding (247 was empirically found to be an effective threshold on average). The blobs surrounded by the resulting contour areas, determined by filling the white areas by the region growing algorithm, are considered as candidate objects in the image. The allowed length of this paper does not allow us

to provide further details on our pre-processing algorithms. However, also other algorithms can be used interchangeably.

3.2 Feature Extraction / Representation

Given a blob, the associated *shape* is a more refined description ready to be compared to the available models (expressed in the form of shapes, as well). Blob border was found to be a very indicative feature for object recognition [8]. Specifically, Fourier descriptors based on distance from the centroid of the shape to its contour points proved to be very effective. Thus, we were inspired by this indicator to set up our approach. Differently from Fourier descriptors, however, we consider the distance from the centroid for contour pixels, and pick just those intersecting selected radian lines at pre-defined angles, originated in the shape centroid. More precisely, a shape is described by a histogram of n sampled values, each normalized to the integer interval $[0 \dots k]$, taken at equally spaced angles from the positive X axis in a coordinate system centered in the blob centroid. The bottom-left of Figure 1 shows a graphical representation of two shapes using both radians (above) and the corresponding 'unrolled' histogram (below).

A first question is how to choose the representation range to represent the single sampled values. Normalizing all the sampled values of a shape to a fixed k ensures scale invariance. We empirically found that a scale of 256 integer values provides a sensible tradeoff between accuracy and tolerance to noise in the blob contours (while requiring just a single byte in memory). So, we associate the centroid to value 0 and the largest sample in a shape to value $k = 255$. Another crucial decision concerns the number n of samples to be taken, in order to have a sufficiently accurate representation without excessively burdening the system. Clearly, the proper tradeoff also depends on the size of the database of models to be matched, and on the kind of objects the system is intended to handle. Next subsections will explain how we set such a parameter, and why. This representation ensures invariance with respect to translation (no information on spacial placement is stored), scale (that does not affect the data structure, but just the values it contains), and intrinsic variables of the images such as luminance and color (completely ignored by the representation, although more refined techniques are to be included in future work). It is also robust to 2D-rotation (by rotating the histogram) and mirroring (by mirroring the histogram).

3.3 Shape Matching

Once the information about the candidate shapes in an image has been extracted, provided a database of sample relevant shapes of interest ('models') is available, the extracted shapes can be compared to those models for possible matching in order to determine the shape class it belongs to. In the bottom row of Figure 1, the left part refers to a query shape extracted from the image, while the center refers to a model shape in the database. Note that each class may be associated to many sample models in the database, and that the matching is performed against the models, not the classes. So, even if the number of classes is kept

constant, increasing the number of models in the database also increases the recognition effort. The expected outcome of the matching between a query shape and a model is a similarity/distance value among the two compared elements. We compare their histograms, representing the distance from the centroid of the blob border in each of the radian directions, according to the intuition that, the more deformed is an object with respect to the model, the more different they are. Specifically, we proceed by overlapping them and summing the absolute pairwise differences of corresponding bars to obtain the overall evaluation (in this case, representing a distance). Another option might be using the statistical measure of variance, but since in our case both the number of values and the values are normalized, a simple summation provides the same results with much less effort. Moreover, for rotation invariance, one such comparison for each displacement of the histogram to be classified over that of the model (considering the histograms as if the last bar were immediately followed by the first one) is needed, displacing each time the histogram by 1 degree to the left, for a total of comparisons equal to the number of bars considered, and then the best case (i.e., the minimum distance value) is taken. The outcome is shown in the bottom-right of Figure 1, where the model shape has been rotated to the best-matching position, and the rightmost histogram shows the pairwise differences among the bars of the shape and model histograms on the left for such an alignment. Overall, if there are n bars to be compared, the effort consists of $c = n \cdot n$ comparisons (subtractions). For mirroring-independence one must double the effort, repeating the above procedure and proceeding in the opposite directions when rotating the histogram (from left to right in one case, and from right to left in the other).

Figure 2 shows the sensitivity of the proposed technique to different geometrical transformations for a sample image (left shape) and corresponding modifications (right shape). For each comparison, the best-matching alignment of histograms is shown, along with the corresponding difference histogram (rightmost histograms). Invariance to translation trivially holds. Invariance to rotation (top case) is proved, since the difference between the shapes is so close to zero that the bars are not visible in the difference histogram. As to scaling (middle case), the difference is visible, but nevertheless small. Also changing the image colors, in this case by considering the negative of the image (bottom case) has a slight effect on the comparison, due to the different outcome of the segmentation step.

3.4 Progressive Approach

Since the matching effort is quadratic in the number of bars to be compared, the basic version of the technique described above might turn out to be inefficient as long as the database size grows. Our solution to tackle this problem consists in a progressive matching procedure, that starts with a few comparisons, and repeatedly selects the most similar models only, to be carried on to a next matching step including more comparisons, until a single model neatly wins or the maximum number of comparisons has been reached.

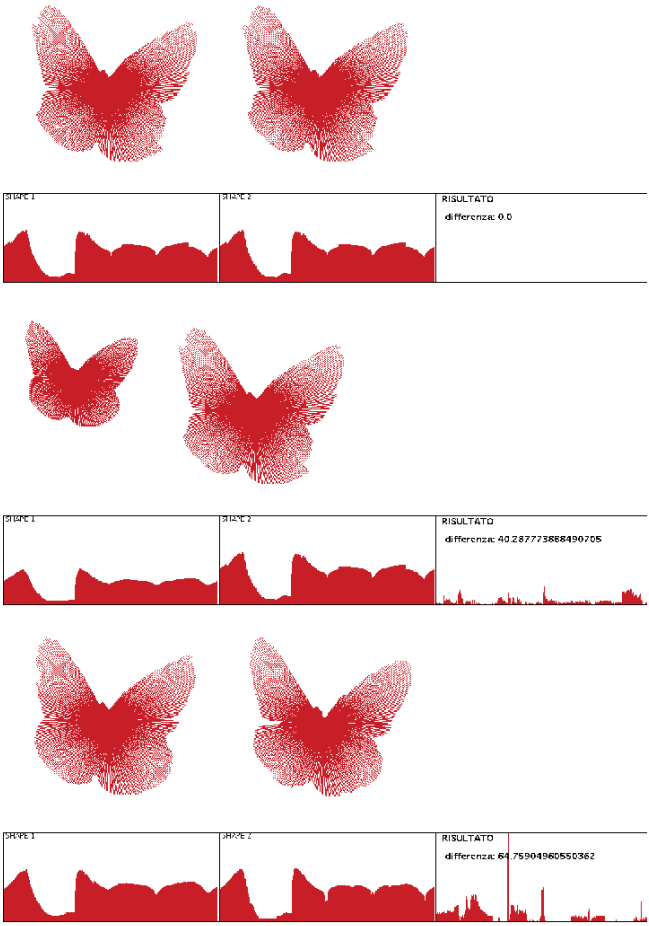


Fig. 2. Check of invariance on a sample image

A simple and straightforward way for increasing the number of comparisons at each matching stage is doubling it, which would make more comfortable the use of powers of 2. an angle displacement of 1.40625 degrees between consecutive radians). In fact, the binary system for angle measurement divides the round angle into 256 degrees, called *brads* (from Binary RADianS). Thus, a straight angle consists of 64 brads, and angles can be comfortably represented using a single byte (more in general, an integer number of bytes — but in our case 2 bytes would be already too much).

The first stage in the matching algorithm compares just 16 values (less comparisons would be too limited to provide a sensible indication on the actual shape), sampled at $16 \cdot i$ brads ($i = 0, \ldots, 15$) along the raw shape, to the 256 values representing a model, for a total of just $16 \cdot 256 = 4096$ comparisons for each shape in the database. Due to the doubled sampling frequency technique,

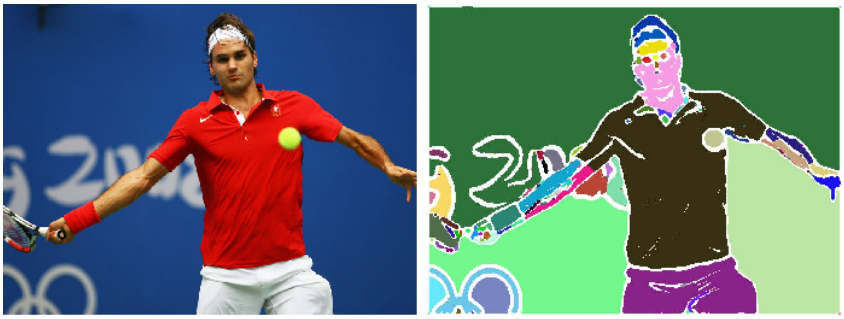


Fig. 3. A picture involving a combination of several shapes

the samples considered at each next step are a superset of those in the previous one, and hence the number of new comparisons per shape is, respectively, 4096, 8192, 16384 and 32768 in the last step.

3.5 Prospects

The final aim of this kind of processing, that we intend to develop as future work, is to be able to understand the content of a whole picture based on the shapes recognized in it and to the particular spatial relationships existing among them. E.g., the picture in Figure 3 might be classified as 'sports' or 'tennis' due to the presence of shirt, shorts, face, ball and racket shapes, where the face is just above the shirt that, in turn, is just above the shorts. To do this, First-Order Logic descriptions are needed, to be able to express relationships among shapes. In the above example, the description might run as:

contains(p,f), face(f), contains(p,t),shirt(t), contains(p,s), shorts(s),
contains(p,b), ball(b), above(f,t), above(t,s), above(f,b), above(b,s), overlap(b,t)

The classification rules might even be learned using ILP systems. In particular, due to the incremental behavior of the shape-learning technique, an incremental ILP technique should be exploited as well, such as those proposed in [3].

4 Evaluation

The preliminary controlled experiments reported in [4] show that the proposed technique is a viable solution, that can efficiently and effectively recognize known shapes in new images. The reported performance refers to a PC endowed with a Dual Core processor at 2GHz and running Windows Vista. A database of 50 selected model shapes was set up, and used to answer pictorial queries consisting of new images (not used to build the database).

In one experiment, the query image contained a single (unknown) shape to be compared against a subset of 16 models chosen at random from the database:

tree (2 models), flower (2 models), shirt, pear, face, star, fish, man, butterfly, fence, moon, sea star, banana, glasses. 3 steps (up to 64 comparisons) were needed before finding a neat winner model for classification. Less than 5 msec were taken for each matching in the first stage, no more than 30 msec for the third stage. Each step discarded all surviving models whose similarity was below the average similarity among the models in the previous step: 5 shapes were discarded in the first step, 3 more in the second one, and finally 7 in the third, which determined the winner.

Table 1. Matching performance for a single shape

Initial models	step 1	step 2	step 3	step 4
16	5/11	3/8	-	-
25	10/15	6/9	-	-
35	14/21	7/14	6/8	-
50	21/29	12/17	7/10	-
Comparisons	16	32	64	128

Then, in the second experiment the subset of models was increased from 16 to 25, then to 35 and finally to 50 shapes. The number of filtered/surviving models at each step for each size is reported in Table 1, showing that the larger the database, the more shapes are cut off at each step. While for the 16- and 25-shape sizes 64 comparisons were sufficient to complete recognition, for databases including 35 and 50 models one more step (128 comparison) is needed. Interestingly, the system never required to run the last step (256 comparisons).

The third experiment evaluated the effort required to process pictures involving several shapes. Specifically, a picture including 5 shapes was used as a query against the same subsets of the model shape database, of increasingly larger size, as in the second experiment. As expected, the time needed to process the whole picture is linear in the number of shapes to be processed (taking about 1 sec per shape). Much more interesting is the fact that the size of the database seems to very marginally affect the effort: the gap in shape recognition time between the 16 models case and the 50 models case is just 0.2 sec, and the time curves for the databases sized 25 and 35 in fact overlap.

Now, we present a novel experiment aimed at assessing the convergence performance of the algorithm for several kinds of shapes having different peculiarities. We considered a set of heterogeneous images including complex shapes in different contexts, positions and orientations, and image composition. The dataset included 250 images of 15 different shapes, as summarized in Table 2 (first and second column). It was collected from various repositories on the Internet[1]. Sample images for each shape type are shown in Figure 4. Specifically, a separate experiment was run for each shape class, and an incremental approach was adopted for each experiment. The models database was initialized with all

[1] The shape dataset can be downloaded at `http://lacam.di.uniba.it/~ferilli/ufficiale/res/shapes_dataset.zip`.

Table 2. Dataset composition

Category	Images	Recognized	Last failure
Red Cross images in various scenarios	35	23 (68%)	35
Human hands	20	13 (68%)	13
Fighter aircrafts	30	18 (62%)	16
Simple geometric shapes in various orientations	10	9 (100%)	–
Trifoil	10	9 (100%)	–
Bat shape	20	17 (89%)	3
Ferrari horse	10	9 (100%)	–
W letter	20	18 (95%)	3
S Letter	20	19 (100%)	–
Machine gun, AK-47	10	7 (78%)	4
Mozart	15	13 (93%)	7
Hen	10	9 (100%)	–
House	20	15 (83%)	5
Key	10	8 (89%)	2
Italy Map	10	9 (100%)	–
Average		88%	

the images in the other classes; then the available images for the class under consideration were progressively submitted to the system, and whenever an image was not correctly recognized, it was added to the database before submitting the next images. We disabled mirroring tolerance in this experiment, to check whether and how it affects recognition performance.

As expected, the recognition performance improves with the growing number of images (i.e., models) in the database (remember that many models may be associated to one class): as more models are available in the repository, the number of recognized images grows constantly. Table 2 shows in the third column the number and percentage of recognized images (denoting accuracy), and in the fourth column the number of the last non-recognized image that causes an addition to the models database (denoting how quick the convergence is). The figures are computed ignoring the first image, that (being a shape not yet present in the repository) is obviously not recognized. In most cases, the initial performance depends on the images already added to the database, which explains why at the beginning some images are not recognized. Then gradually, by enriching the database incrementally, the sequence of images that are correctly recognized grows steadily. Of course, even after many images have been added to the database, there may be cases that are not correctly recognized and this is due to the peculiarities of these images. For instance, in the 'Red Cross' class the last image is not recognized. However, in most cases normal shapes already inserted in the database are correctly identified.

As expected, the most problematic cases are those in which the target object is taken in different perspectives and contexts, which results in significant changes in their possible shapes. Indeed, for the 'Red Cross', 'Hands' and 'Aircrafts' subsets, looking at the specific images not recognized one can see that they

Fig. 4. Sample shapes from the test dataset

differ from the already inserted images mainly in their orientation and in the configuration of the background. Conversely, for more standard shapes, such as 'Geometric shapes', 'Trifoil', 'Ferrari', letter S and Italy, the addition of the very first shape is sufficient to completely and correctly learn its model. Also, the good performance on letters might indicate that the proposed technique can be a valuable tool for recognizing printed symbols in general. This raises an additional question, concerning how many images the technique needs in the database in order to have a stable performance. Differently from the cases just discussed, where just one tagged image is sufficient for the technique to subsequently recognize that shape in all future images, for some other shapes it takes more images to make the technique start recognizing shapes steadily. From our observations we found that, in most cases, it is necessary and sufficient to insert in the database images that represent distinguishing features of the shape orientation or direction. For example, for the Mozart silhouette, the two images that were added to the model database are those depicting Mozart in left and right orientation, respectively. After the latter was added to the database, the system recognized successfully all the following images that represent both left and right orientation. This confirms that the shape orientation is important in order for the system to behave correctly, and that in some cases mirroring tolerance is necessary. It is a topic of future work assessing the tradeoff between accuracy and computational cost due to the introduction of mirroring tolerance.

5 Conclusion

Information expressed by images can be hardly accessed, due to the *semantic gap* separating the raw set of pixels from their overall perceptual meaning. Nevertheless, images are very information-dense elements, and hence being able to understand their content would help to improve image indexing and retrieval in digital libraries. This work specifically focuses on Object Recognition, as a fundamental task towards a high-level description of the image content in terms of the objects contained and their inter-relationships. A progressive technique is proposed, that integrates and improves a set of existing representation and processing techniques for identifying objects belonging to known classes for which model shapes are available. A prototype implementation of the proposed approach suggests that effective recognition can take place, with reasonable efficiency in terms of time and space resources. It can recognize objects based on their shape, independently of scaling, translation, mirroring and (2D) rotation.

Future work will concern finding a mix of features that are sufficiently complementary to significantly improve recognition performance over application of shape recognition alone, while not increasing excessively the computational burden. Moreover, we are working on devising strategies for exploitation of the high-level description provided by this technique, both for document understanding and indexing. Other directions for investigation concern the improvement of the pre-processing step, for providing a better input to the recognition engine, a larger evaluation on benchmark datasets and an assessment of what kinds of images can profitably make use of this technique.

References

1. Brause, R., Arlt, B., Tratar, E.: Project semacode: A scale-invariant object recognition system for content-based queries in images databases. Technical Report 11/99 (FB20), Johann Wolfgang Goethe University, Computer Science Dept., Frankfurt/Main (1999)
2. Chen, Y., Li, J., Wang, J.Z.: Machine Learning and Statistical Modeling Approaches to Image Retrieval. Information Retrieval, vol. 14. Kluwer (2004)
3. Ferilli, S., Basile, T.M.A., Biba, M., Di Mauro, N., Esposito, F.: A general similarity framework for horn clause logic. Fundamenta Informaticae 90, 43–66 (2009)
4. Ferilli, S., Basile, T.M.A., Esposito, F., Biba, M.: A contour-based progressive technique for shape recognition. In: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR 2011), vol. 1, pp. 723–727. IEEE Computer Society (2011)
5. Hogendoorn, H.: The state of the art in visual object recognition (2006)
6. Shu, X., Wu, X.-J.: A novel contour descriptor for 2d shape matching and its application to image retrieval. Image and Vision Computing 29(4), 286–294 (2011)
7. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer (2011)
8. Zhang, D., Lu, G.: A comparative study of curvature scale space and fourier descriptors. Journal of Visual Communication and Image Representation 14(1), 41–60 (2003)

EDB: Knowledge Technologies for Ancient Greek and Latin Epigraphy

Fabio Fumarola[1], Gianvito Pio[1], Antonio E. Felle[2], Donato Malerba[1], and Michelangelo Ceci[1]

[1] Dipartimento di Informatica, Università degli Studi di Bari "A. Moro"
via Orabona, 4 - 70126 Bari, Italy
{fabio.fumarola,gianvito.pio,donato.malerba,michelangelo.ceci}@uniba.it
[2] Dipartimento di Scienze dell'Antichità e del Tardo Antico,
Università degli Studi di Bari "A. Moro"
strada Torretta (Città Vecchia) - 70122 Bari, Italy
antonio.felle@uniba.it

Abstract. Classical Greek and Latin culture is the very foundation of the identity of modern Europe. Today, a variety of modern subjects and disciplines have their roots in the classical world: from philosophy to architecture, from geometry to law. However, only a small fraction of the total production of texts from ancient Greece and Rome has survived up to the present days, leaving many ample gaps in the historiographic records. Epigraphy, which is the study of inscriptions (epigraphs), aims at plug this gap. In particular, the goal of Epigraphy is to clarify the meanings of epigraphs, classifying their uses according to dates and cultural contexts, and drawing conclusions about the writing and the writers. Indeed, they are a kind of cultural heritage for which several research projects have recently been promoted for the purposes of preservation, storage, indexing and on-line usage. In this paper, we describe the system EDB (Epigraphic Database Bari) which stores about 30,000 Christian inscriptions of Rome, including those published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* editions. EDB provides, in addition to the possibility of storing metadata, the possibility of *i)* supporting information retrieval through a thesaurus-based query engine, *ii)* supporting time-based analysis of epigraphs in order to detect and represent novelties, and *iii)* geo-referencing epigraphs by exploiting a spatial database.

Keywords: Epigraphy, Information Retrieval, Knowledge Bases, Novelty Detection, Spatial Databases.

1 Introduction

Many countries are nowadays interested in the valorization of the cultural heritage, since it is widely recognized that cultural heritage resources have significant implications for development (both as a knowledge basis and in terms of commercial exploitation). For this aim, many institutions which collect and

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 29–40, 2014.

Fig. 1. An example of epigraph

preserve cultural heritage have shown a great interest in the digitalization of their resources and in the exploitation of mechanisms to provide online access to digitalized products.

According to the definition reported in the 1972 UNESCO "World Heritage Convention" - Article 1 - cultural heritage refers to "monuments", "groups of buildings" and "sites" which are of outstanding universal value historically, artistically or scientifically. However, the concept of cultural heritage has recently assumed a broader connotation and includes, among other things, tangible, moveable objects such as works of art, artifacts, scientific specimens, photographs, books, manuscripts and recorded moving image and sound [2].

In the literature, several systems have been proposed for the analysis, also through knowledge technologies, of digital Cultural Heritage resources and metadata. Typically, they are developed in the context of research projects such as MASTER [13], MEMORIAL [4], D-SCRIBE [11][12], CULTURA [1], PROMISE [10], COLLATE [9] and CDLI (`http://cdli.ucla.edu`). They are either general purpose projects or focused on specific types of artifacts (e.g. COLLATE focuses on digitalized film archives of the Second World War) but none of them take into account the specific case of epigraphs.

Epigraphs are invaluable sources of information that provide us with a myriad of useful information of the past (an example of epigraph is shown in Figure 1). They play the role of "time capsules" for example by allowing us to shed light on otherwise undocumented historical events, or to gain new knowledge of local laws and customs, and even to determine the date and producer of a given piece of lead piping. Epigraphy also documents the evolution of languages and scripts, although indirectly. In some cases, such as that of the Rosetta Stone, it can provide those key insights that allow for the successful deciphering of an unknown script.

In recent years, the tendency to create epigraphic databases for storing both images of epigraphs and associated metadata, as well as for supporting retrieval functionalities has emerged [8]. Examples are [6] and, more recently, [3] which exploit the EPIDOC schema [5] to guarantee uniform representation of epigraphic metadata. In the particular case of [3], the authors propose the *Hispania*

Epigraphica database which allows the representation and the exploitation of semantic links between entities.

However, four main problems have, up to now, affected epigraphic databases:

1. Retrieval capabilities should take into account possible evolutions of the language, possibly due to the influence of other languages (e.g. in the Middle age) which lead to *aberrant* forms. Preserving retrieval effectiveness in these cases requires the consideration of a thesaurus which plays the role of background knowledge to be used for retrieval purposes.
2. Classic epigraphy has evolved into three strictly separate disciplines, i.e. Greek, Latin and Christian epigraphy, characterized by separate collections and corpuses used for reference, separate publications, separate populations of scholars and researchers. Indeed, different languages require different background knowledge to be exploited.
3. Data available in the databases can be used to extract knowledge through the application of data mining algorithms (following previous studies that use data mining algorithms for the analysis of digitalized cultural heritage resources [7]).
4. Epigraphs have traditionally been featured in non-geographical data bases. This results in the fact that several inscriptions that should be linked because of thematic or historic commonalities are scattered across multiple collections. Geo-referencing would help to overcome this limitation.

In this paper, we present EDB (Epigraphic Database Bari) which concentrates on the valorization of the huge Italian cultural heritage and, in particular, on the valorization of Christian inscriptions in Rome. EDB actually stores around 30,000 Christian inscriptions and provides an answer to the problems described before. It stores metadata, such as the type of support (e.g. marble), the approximate period, the engrave technique, the current position and the text. Moreover, in the retrieval phase it expands queries by exploiting a thesaurus which includes more than 3000 relationships between terms. It also integrates a data mining algorithm which faces a novelty detection task. In this way, it is possible to identify relevant (frequent) changes in the properties of the inscriptions over time. Finally, it embeds a spatial database used to geo-reference epigraphs.

The paper is organized as follows. In the next section, we describe the system architecture of EDB. In Section 3, we describe the application of EDB to the inscriptions published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* editions. Finally, in Section 4, we report conclusions and delineate some future work.

2 System Architecture

The general architecture of the proposed system consists of several components, each of which is in charge of performing specific tasks, and of a set of different data sources (see Figure 2). A summarized description of each of them is reported in the following subsections.

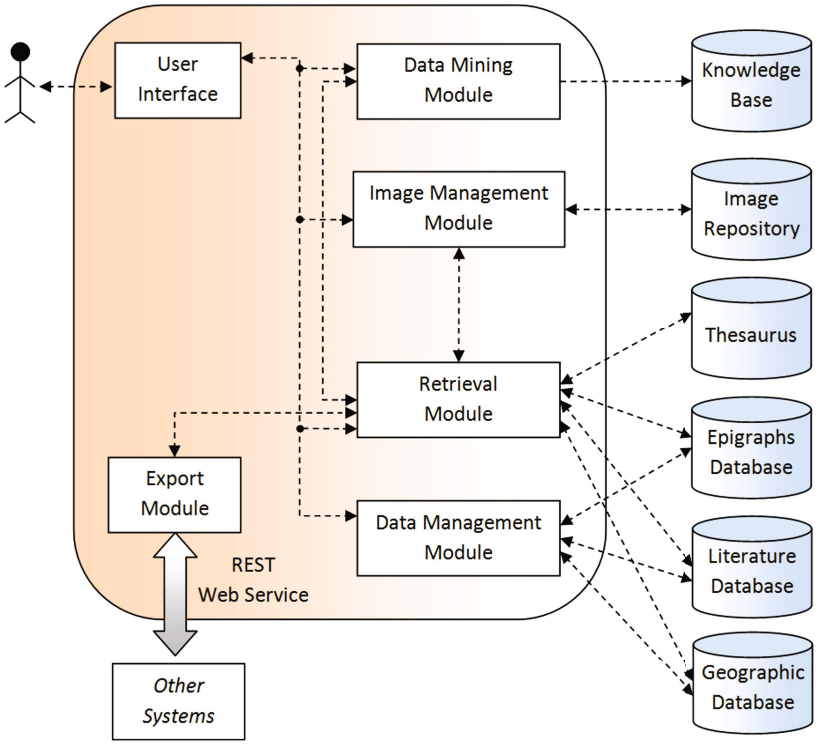


Fig. 2. System architecture

2.1 Data Sources

Since the main goal of the system is to store and retrieve data about epigraphs, the main data source is the *Epigraph Database*, which, together with the text of the epigraphs, stores a set of metadata about dating, original context, current location, related literature, etc.

The *Literature Database* stores metadata about scientific papers in which epigraphs have been studied. These metadata can include, but are not limited to, the authors, the journal, the year of publication, the citations, etc. Each epigraph stored in the Epigraph Database can be associated to one or more scientific papers stored in this database.

The *Knowledge Base* stores the knowledge extracted by the Data Mining Module, i.e. patterns of interest in the form of rules, clusters, etc.

The *Image Repository* stores photos of the epigraphs. This repository can be either internal or external, as well as an integration of internally produced and external resources.

The *Thesaurus* data source is useful to support retrieval tasks. In particular, it gives the possibility to enhance and/or expand the user's query to better match data stored in the epigraph database. This aspect is detailed later in the Section 2.3.

The *Geographic Database* stores data about geographic positions. It can be either an internal or an external resource (e.g. a web service). One or more geographic positions can be associated to each epigraph, which can represent either the locations where the epigraph (or an its fragment) was found or the position where the epigraph is currently located.

2.2 System Components

The *User Interface* component allows users to interact with the system. In particular, this component is in charge of translating each user action to a request to other system components and to properly show the returned results. Although in Figure 2 a single actor is reported, it is noteworthy that different types of users may interact with the system. In particular:

- **Administrators**, which access the system to manage users and services;
- **Compilers/Epigraphists**, which are the domain experts in charge of inserting, editing and deleting data about epigraphs to/in/from the main database, as well as of managing the image repository and the literature related to the epigraphs;
- **Web users**, which are mainly interested in retrieving information about archived epigraphs according to many different filtering criteria;
- **Data Analysts**, whose goal is to analyze data in order to extract valuable knowledge from them.

The *Retrieval Module* exposes an interface to query the Epigraph Database. This is the central component of the system, through which users, as well as other components of the system, can access data to perform their own tasks.

In particular, the main goal of this component is to retrieve epigraphs which satisfy a given set of filtering criteria. Such filters can be defined on the epigraphs' text, as well as on their metadata, which include the dating, the related literature and the geographic position about their original context or about their current location. Therefore, this component has to access directly to the Epigraph Database, to the Literature Database and to the Geographic Database. Furthermore, since performing the retrieval on the basis of the epigraph's text could be tricky when aberrant forms are present (see Section 2.3), this component also exploits the Thesaurus data source.

The *Export Module* offers a service to other systems that need to access to information about epigraphs, acting as a bridge between them and the Retrieval Module.

The *Data Management Module* performs the tasks which are mainly related to the activity of compilers. In particular, this component allows the users to insert, edit and delete data about the epigraphs and manage all their aspects, such as the related literature and the geographic positions associated to the original context or to the current location of conservation.

The *Image Management Module* allows the users (mainly compilers) to manage, i.e. insert into, editing or deleting from, the image repository. Part of the

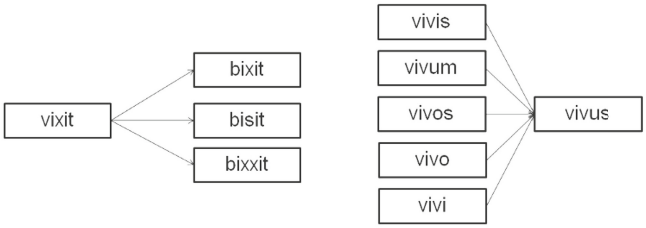


Fig. 3. On the left, an example of aberrant forms of the term "vixit". On the right, a set of terms which can be mapped to the same term "vivus".

image repository is obtained via web queries to the *Pontificia Commissione di Archeleogia sacra* (PCAS) web site[1].

The *Data Mining Module* allows the users (mainly data analysts) to execute data mining algorithms on the available data, in order to discover valuable knowledge, which is then stored in the Knowledge Base. This component accesses data through the Retrieval Module. An example of data mining task which can be applied to epigraphs is reported in Section 2.4.

2.3 Dealing with Aberrant Forms through a Thesaurus

Among all the possibilities that a user can exploit to find an epigraph of interest, the text-based search on the inscription is one of the most straightforward way provided by the Retrieval Module.

However, in this particular domain, a simple matching strategy between the query and the text of the inscriptions stored in the database can easily fail, since *i)* the same term could have changed in its phonetic or orthography over time (aberrant forms) and *ii)* many different terms semantically map to the same concept (see Figure 3).

In this context, it is useful to design a thesaurus containing the relations between aberrant forms and normal forms as well as the sets of terms which map to the same concepts. In this way, the quality of the results returned by the Retrieval Module can be substantially improved, since it can easily map each term contained in the query and in text of the inscriptions to its normal form (and/or to the term which express the general concept), before verifying the matching between the query and the inscription.

2.4 Novelty Detection

In this subsection we report one of the possible data mining tasks, called *novelty detection*, which can be applied on data about epigraphs. In particular, the main goal of this task is to identify emerging patterns, that is, patterns which show relevant changes (in frequency) over time. Therefore, this task is strongly related to the temporal dimension associated to the epigraphs.

[1] `http://pcas.xdams.net/pcas-web/home.html`

Since epigraphs can usually be associated to an historical period (dating), even to a single year or to a definite time interval, it could be interesting to identify how social and cultural changes over time have affected the epigraphs, in the phonetics or in the orthography as well as in the used materials or in the executing techniques.

This task requires to deal with several pre-processing issues. In particular:

- Identification of the *reference objects* and of the *task-relevant objects*. In this case, target objects are clearly the epigraphs, while the task relevant objects are the materials, the executing techniques, the different kinds of writing, etc.
- Definition of proper time intervals (or time windows), which consists in the identification of an adequate number of intervals and in the choice of the discretization method to apply (e.g. equal width, equal frequency, clustering-based discretization).
- Feature selection, that is, identification of features of interest among all the available ones, in order to focus the algorithm only on the relevant data.

In the following we report two examples of possible emerging patterns, expressed as a list of logical predicates, which could be discovered by applying novelty detection algorithms to data about epigraphs.

$$[250, 349] \rightarrow [350, 399] : epigraph(E), transcription(E, T), term(T, \text{“vixit”}) \quad (1)$$

$$[250, 349] \rightarrow [350, 399] : epigraph(E), pertinence\_area(E, \text{“Via Appia”}) \quad (2)$$

In the example (1), the discovered pattern describes a relevant increase of the number of epigraphs containing the term “vixit” in their transcription, in the time interval [350, 399] with respect to the time interval [250, 349]. The pattern in the example (2) emphasizes an increased amount of epigraphs in the area of “Via Appia”, in the same period.

The discovered patterns can be ranked according to some measures of relevance, such as the *growth rate*, which represents the relative variation of the support of the pattern in the considered time intervals.

It is noteworthy that the discovered patterns can suggest the researchers some relevant aspects that are worth to be deeply investigated, since they can describe social and cultural changes otherwise difficult to identify in the huge amount of available data.

3 Application to Roman Inscriptions

EDB (Epigraphic Database Bari)[2] is an online, freely accessible, database hosted by the University of Bari. It includes about 30000 Christian inscriptions of Rome,

[2] `http://www.edb.uniba.it`

including inscriptions published in the *Inscriptiones Christianae Vrbis Romae septimo saeculo antiquiores, nova series* (ICVR) editions. The *ICVR* editions started in 1922 with the first volume and is going to end with the eleventh volume in the next years.

Similar initiatives are EDR (Epigraphic Database Roma)[3], EDH (Epigraphic Database Heidelbergh)[4] and HE (Hispania Epigraphica)[5]. However, they actually do not offer the possibility of performing complex text-based search, also with the help of the thesaurus. Furthermore, data about epigraphs are not entirely reported in each of these databases, since they may focus on different specific aspects. The project EAGLE (Europeana network of Ancient Greek and Latin Epigraphy), indeed, aims at the creation of a federation of these databases (including EDB), in order to allow the researcher to retrieve data about epigraphs in an integrated way from all the databases.

Currently, EDB includes the text and the metadata of around 30000 inscriptions, discovered, classified and enriched with semantic metadata by Carlo Carletti and his team. EDB makes freely available over the web the inscriptions

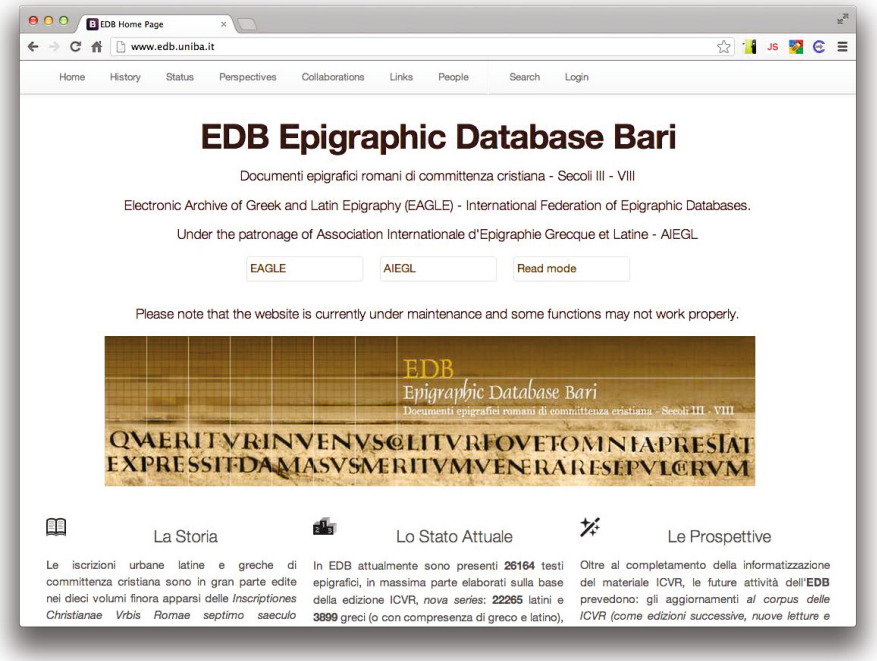
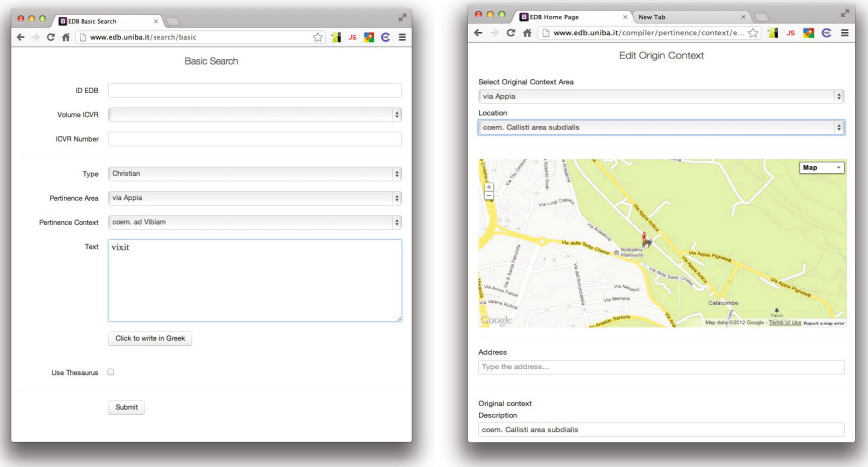


Fig. 4. Screenshot of the EDB's main page

[3] http://www.edr-edr.it
[4] http://edh-www.adw.uni-heidelberg.de
[5] http://eda-bea.es



(a) Screenshot of the search web page. (b) Screenshot of the geographic position for the context *coem. Callisti area subdialis*.

Fig. 5. Examples of functional screenshots of the EDB web interface

discovered over 25 years of archeological studies inside the Roma's catacombs. EDB allows access to an unique cultural heritage which is in its major part not accessible by public visitors.

Figure 4 shows the main page of the EDB web site. In order to maximize the user experience, EDB is implemented using HTML5 and CSS3. This allows users to fully access the information stored in EDB using a common web browser, a tablet or a smartphone. Moreover, all the functionalities of the web applications are exposed through a *restful* interface [14].

The main functionality offered by EDB is in the epigraph search. In Figure 5a the basic search web page is presented. Epigraphs can be retrieved on the basis of their EDB identifier, their ICVR volume and number, by the religious identity of the epigraph (Christian, Jewish and Pagan), by the area and context of pertinence, as well as by specifying a textual query and using the thesaurus. The text for a query can be written using latin words and greek words by enabling an automatic greek inputter inserted in the query web page. For example (see Figure 5a) if the user searches for epigraphs of type *Christian*, which are discovered on the *via Appia*, in the context *coem. ad Vibiam* and such that their text contains *vixit*, EDB currently retrieves 11 matching inscriptions (see Figure 6). This figure shows the text of the inscriptions and the related metadata.

Fig. 6. Screenshot of a query result page of EDB

Figure 7 shows information stored for the epigraph *14387*. In particular, it shows information on (from the top left corner) the ICVR volume, the reference literature, the area of pertinence and its context, the place where the inscription is physically stored, the support and the technique used, the editing time (which is estimated by archeologists), a link to the image of the inscription on the PCAS web site and, finally, its text.

Moreover, thanks to the spatial database, all the inscriptions stored in EDB are geo-referenced. Figure 5b shows as example the geographic position of the *coem. Callisti area subdialis* situated in the *via Appia*.

All the stored information allow us to evaluate the effort made by the archeologist and by EDB ecosystem to add valuable information to the text of the epigraph. This is the added value of the EDB project.

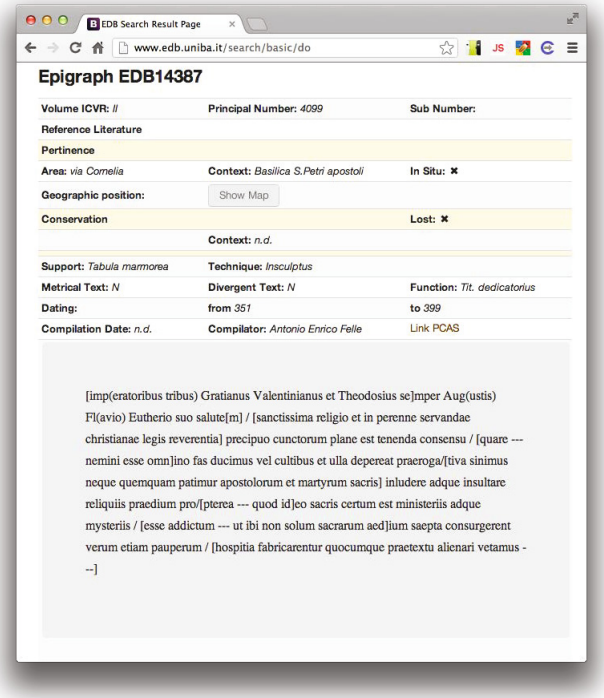


Fig. 7. Screenshot of the epigraph EDB-14387

4 Conclusions

This paper presents the system EDB which supports epigraphists in storing, managing and retrieving information on a large repository of Italian inscriptions found in the area of Rome. Peculiarities of the proposed system are in the possibility of managing, in addition to metadata, geo-spatial information and references which cite the specific epigraph. EDB also supports time-based analysis of epigraphs which aims at detecting and representing changes of inscriptions in their properties/metadata over time.

By means of its web interface, it also allows (possibly non-expert) web-users to define search queries and retrieve all the necessary information. Search queries are automatically expanded by the system, according to a thesaurus, in order to consider aberrant forms.

The effectiveness of EDB is proved by its current and extensive use by a team of epigraphists which elaborate a set inscriptions discovered over 25 years of archeological research and studies inside the Roma's catacombs.

Acknowledgments. The authors thank Carlo Carletti, first scientific responsible of the EDB project since 1988, Antonella Daniela Agostinelli and Anita Rocco as well as past and present collaborators: Cristina Grisanzio, Ruggero

Lombardi, Filippo Piazzolla, Marida Pierno, Miriam Ramunni, Domenico Schiraldi and Carolina Ventura.

This work partially fulfills the research objectives of the PON 02_00563_3470993 project "VINCENTE - A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" funded by the Italian Ministry of University and Research (MIUR).

References

1. Agosti, M., Benfante, L., Orio, N.: A contribution for the dissemination of cultural heritage content to a wider public. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 195–206. Springer, Heidelberg (2013)
2. Altamura, O., Berardi, M., Ceci, M., Malerba, D., Varlaro, A.: Using colour information to understand censorship cards of film archives. International Journal on Document Analysis and Recognition 9(2-4), 281–297 (2007)
3. Álvarez, F.-L., Gómez-Pantoja, J.-L., Barriocanal, E.G.: From relational databases to linked data in epigraphy: Hispania epigraphica online. In: Barriocanal, et al. (eds.) [5], pp. 225–233
4. Antonacopoulos, A., Karatzas, D.: Document image analysis for World War II personal records. In: 1st International Workshop on Document Image Analysis for Libraries (DIAL), pp. 336–341 (2004)
5. García-Barriocanal, E., Cebeci, Z., Okur, M.C., Öztürk, A. (eds.): MTSR 2011. CCIS, vol. 240. Springer, Heidelberg (2011)
6. Bodard, G.: The inscriptions of aphrodisias as electronic publication: A user's perspective and a proposed paradigm. Digital Medievalist 4 (2008)
7. Ceci, M., Berardi, M., Malerba, D.: Relational data mining and ilp for document image understanding. Applied Artificial Intelligence 21(4&5), 317–342 (2007)
8. Feraudi-Gruénais, F.: Latin on Stone: Epigraphy and Databases. Lexington Book (2010)
9. Frommholz, I., Brocks, H., Thiel, U., Neuhold, E.J., Iannone, L., Semeraro, G., Berardi, M., Ceci, M.: Document-centered collaboration for scholars in the humanities – the COLLATE system. In: Koch, T., Sølvberg, I.T. (eds.) ECDL 2003. LNCS, vol. 2769, pp. 434–445. Springer, Heidelberg (2003)
10. Gäde, M., Ferro, N., Paramita, M.L.: CHiC 2011 - Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. In: Petras, V., Forner, P., Clough, P.D. (eds.) CLEF (Notebook Papers/Labs/Workshop) (2011)
11. Gatos, B., Ntzios, K., Pratikakis, I., Petridis, S., Konidaris, T., Perantonis, S.J.: A segmentation-free recognition technique to assist old greek handwritten manuscript OCR. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 63–74. Springer, Heidelberg (2004)
12. Gatos, B., Pratikakis, I., Perantonis, S.J.: An adaptive binarization technique for low quality historical documents. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 102–113. Springer, Heidelberg (2004)
13. Le Bourgeois, F., Kaileh, H.: Automatic metadata retrieval from ancient manuscripts. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 75–89. Springer, Heidelberg (2004)
14. Richardson, L., Ruby, S.: RESTful web services. O'Reilly Media, Incorporated (2007)

Fostering Interaction with Cultural Heritage Material via Annotations: The FAST-CAT Way

Nicola Ferro[1], Gary Munnelly[2], Cormac Hampson[2], and Owen Conlan[2]

[1] Department of Information Engineering, University of Padua, Italy
ferro@dei.unipd.it
[2] The University of Dublin, Trinity College, Ireland
{munnelg,cormac.hampson,owen.conlan}@scss.tcd.ie

Abstract. This paper describes the innovative annotation facilities of the CULTURA portal for digital humaties, which are aimed at improving the interaction of non specialist users and general public with cultural heritage contents. The annotation facilities are comprised by two modules: the FAST annotation service as back-end and the CAT Web front-end integrated in the CULTURA portal.

1 Introduction

Almost everybody is familiar with annotations and has his own intuitive idea about what they are, drawn from personal experience and the habit of dealing with some kind of annotation in everyday life, which ranges from jottings for the shopping to taking notes during a lecture or even adding a commentary to a text. This intuitiveness makes annotations especially appealing for both researchers and users: the former propose annotations as an easy understandable way of performing user tasks, while the latter feel annotations to be a familiar tool for carrying out their own tasks. Therefore, annotations have been adopted in a variety of different contexts, such as content enrichment, data curation, collaborative and learning applications, and social networks, as well as in various information management systems, such as the Web (semantic and not), digital libraries, and databases.

The role of annotations in digital humanities is well known and documented [1-6]. Subsequently, many different tools which allow for the annotation of digital humanities content have been developed. Unfortunately, tools designed specifically for an individual portal are typically only compatible with that system. More general solutions, which can be easily distributed across various sites, have been developed, but these systems often have limited functionality (only annotating a single content type, no sharing features etc.) [7-8].

FAST-CAT (Flexible Annotation Semantic Tool - Content Annotation Tool) is a generic annotation system that directly addresses this challenge by providing a convenient and powerful means of annotating digital content. This paper introduces FAST, the backend service providing powerful annotation functionalities, and CAT, the frontend Web annotation tool, and discusses how its features are tackling important challenges within the Digital Humanities field.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 41–52, 2014.

FAST-CAT is being developed as part of the CULTURA project [9-10]. A key aspect of CULTURA is the production of an online environment that empowers users, of various levels of expertise, to investigate, comprehend and contribute to digital cultural collections. FAST-CAT is a key component of this environment and is currently being trialed with the help of three different user groups.

The paper is organized as follows: Section 2 explains the fast annotation model and the search functionalities on top of it; Section 3 describes the CAT annotation interaction model; Section 4 introduces the FAST-CAT architecture; Section 5 discusses the CULTURA environment; and, Section 6 draws some conclusions and outlook future work.

2 FAST Annotation Model

The FAST annotation service adopts and implements the formal model for annotations proposed by [3] which has been also embedded in the reference model for digital libraries developed by DELOS, the European network of excellence on digital libraries [11].

According to this model, an annotation is a compound multimedia object which is constituted by different signs of annotation. Each sign materializes part of the annotation itself; for example, we can have textual signs, which contain the textual content of the annotation, image signs, if the annotation is made up of images, and so on. In turn, each sign is characterized by one or more meanings of annotation, which specify the semantics of the sign; for example, we can have a sign whose meaning corresponds to the title field in the Dublin Core (DC) metadata schema, in the case of a metadata annotation, or we can have a sign carrying a question of the author's about a document whose meaning may be "question" or similar.

An annotation has a scope which defines its visibility (public, shared, or private), and can be shared with different groups of users. Public annotations can be read by everyone and modified only by their owner; shared annotations can be modified by their owner and accessed by the specified list of groups with the given access permissions, e.g. read only or read/write; private annotations can be read and modified only by their owner.

Figure 1 shows an example of annotation which summarizes the discussion so far. The annotation, with identifier `a1` and namespace `fast`, is authored by the user `ferro`. It annotates a document containing a novel, whose identifier is `doc1` and which belongs to the namespace `dl1` of a digital library which manages it. The annotation relates to another document containing a translation of the novel, whose identifier is `doc35` and which belongs to the namespace `dl2` of a digital library different from the one which manages `doc1`; in addition, it relates also to the Web page of the publisher of the novel, whose identifier is `http://www.publisher.com/` and which belongs to the namespace `fweb`, used for indicating Web resources.

In particular, `a1` annotates two distinct parts of `doc1`. It annotates an image contained in the PDF of the novel by using a textual sign whose content is "This is a common picture for this novel" and whose meaning is to be a `comment` in the `fast` namespace. It also

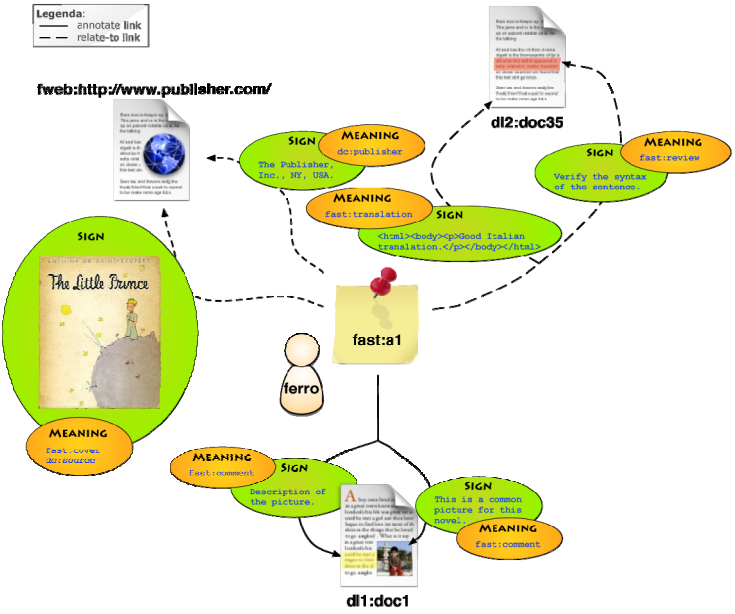


Fig. 1. Example of annotation

annotates a sentence by using another textual sign whose content is "Description of the picture" and whose meaning is to be a `comment` in the `fast` namespace.

`a1` relates the document doc1 to its Italian translation by linking to the whole document `doc35` with a textual sign whose content is "Good Italian translation" and whose meaning is to be a `translation` in the `fast` namespace. It also relates to a specific sentence of the translation with a HTML sign which asks to "Verify the syntax of the sentence" and whose meaning is to be a `review` in the `fast` namespace. Finally, `a1` also relates the document to the Web page of the publisher of the novel with a textual sign whose content is "The Publisher, Inc., NY, USA" and whose meaning is to be the `publisher` field of the `DC` metadata schema. It also relates the document to the Web page of the publisher via an image sign, containing the cover of the printed book of the novel by the publisher, and whose meaning is to be both a `source` field in the `DC` metadata schema and a `cover` in the `fast` namespace.

The flexibility inherent in the annotation model allows us to create a connective structure, which is superimposed to the underlying documents managed by digital libraries. This can span and cross the boundaries of different digital libraries and the Web, allowing the users to create new paths and connections among resources at a global scale.

2.1 Search Model

The presence of both structured and unstructured content within the managed resources calls for different types of search functionalities, since structured content can

be dealt with exact match searches while unstructured content can be dealt with best match searches. These two different types searches may need to be merged together in a query if, for example, the user wants to retrieve annotations by a given author about a given topic; this could be expressed by a boolean AND query which specifies both the author (structured part) and the content (unstructured part) of the annotations to be searched. Nevertheless, boolean searches are best suited for dealing with exact match searches and they need to be somewhat extended to also deal with best match searches. Therefore, we need to envision a search strategy able to express complex conditions that involve both exact and best match searches. The “P-norm" extended boolean model proposed by [12] is capable of dealing with and mixing both exact and best match queries, since it is an intermediate between the traditional boolean way of processing queries and the vector space processing model. Indeed, on the one hand, the P-norm model preserves the query structure inherent in the traditional boolean model by distinguishing among different boolean operators (and, or, not); on the other hand, it allows us to retrieve items that would not be retrieved by the traditional boolean model due to its strictness, and to rank them in decreasing order of query-document similarity. Moreover, the P-norm model is able to express queries that range from pure boolean queries to pure vector-space queries, thus offering great flexibility to the user.

The hypertext that connects documents to annotations calls for a search strategy that takes it into consideration and allows us to modify the score of annotations and/or documents according to the paths in the hypertext. For example, we could consider that an annotation, retrieved in response to a user query, is more relevant if it is part of a thread where other annotations have also been retrieved in response to the same query rather than if it is part of a thread where it is the only annotation that matches the query.

The FAST Context Set [13] has been defined in order to provide a uniform query syntax to FAST by using the Contextual Query Language (CQL) [14], developed and maintained by the Library of Congress in the context of the Z39.50 Next Generation (ZING) project. FAST provides conformance to CQL up to Level 2.

3 CAT Annotation Interaction Model

CAT is a web annotation tool developed with the goal of being able to annotate multiple types of documents and assist collaboration in the field of digital humanities. At present, CAT allows for the annotation of both text and images. The current granularity for annotation of text is at the level of the letter. For image annotations, the granularity is at the level of the pixel. This allows for extremely precise document annotation, which is very relevant to the Digital Humanities domain due to the variety of different assets that prevail. How this precision was achieved is discussed in section 3.1.

There are two types of annotation which may be created using CAT; a targeted annotation and a note. A targeted annotation is a comment which is associated with a specific part of a document. This may be a paragraph, a picture or an individual word, but the defining feature is that the text is directly associated with a specific subset of the digital resource. Conversely, a note is simply attached to the document. It is not

associated with a specific item therein. Typically, this serves as a general comment or remark about the document as a whole.

In addition to allowing a user to comment on document text, the annotations created using CAT allow an individual to link their annotations to other, external sources. This is hugely beneficial for teachers using digital cultural collections and for students from primary to university level as well as experienced researchers. As can be seen in Figure 1, the addition of links to a resource greatly enriches the amount of information it contains. Importantly, each link has comment text associated with it, allowing an educator to explain why this specific link is important or what the student should seek to gain from reading this particular source.

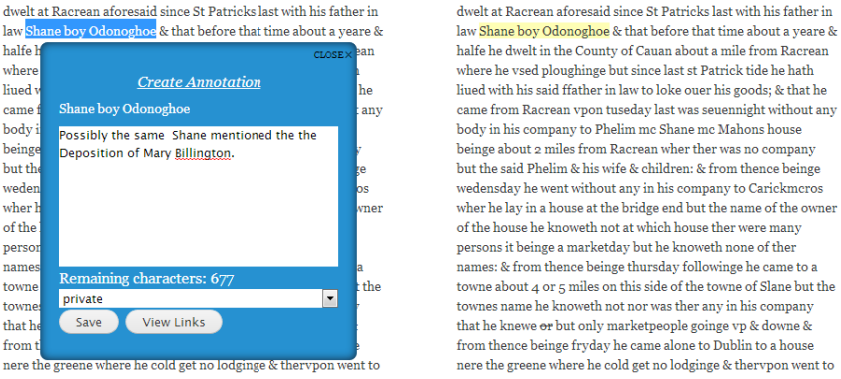


Fig. 2. User creates a targeted annotation on a body of text about a person of interest

While CAT is beneficial for researchers and educators, it is also being used as an important source of user data for the content provider. Websites such as Amazon[1] and YouTube[2] are able to provide increasingly accurate recommendations for their individual users. These recommendations are facilitated by a user model which is driven by a combination of factors such as ratings and recently viewed items. For a digital humanities site, annotations can provide an insight into which entities are of interest to a user. If a user is frequently annotating a document, it is likely that this document is of interest to them. Furthermore, if the text being annotated is analysed, it may be possible to discern specific entities of interest within the document. A digital humanities site which could recommend resources that are relevant to users would be profoundly useful, and would help improve the effectiveness with which researchers interact with their domain.

3.1 Annotation Pointers

Anchoring annotations to the annotated objects is fundamental for CAT to provide the desired degree of annotation precision. This is achieved by serializing a representation of a pointer to the annotated digital object.

[1] http://www.amazon.co.uk/
[2] http://www.youtube.com/

For text, this serialized representation takes the form:

```
<PathStart>;<OffsetStart>;<PathEnd>;<OffsetEnd>
```

Where:

- `<PathStart>` is the path to the element which contains the start of the user's selection.
- `<OffsetStart>` is the offset into the start element where the beginning of the selected text may be found.
- `<PathEnd>` is the path to the element which contains the end of the user's selection.
- `<OffsetEnd>` is the offset into the end element where the end of the selected text may be found.

For images, the form is:

```
<Path>;<OffsetX>;<OffsetY>;<AnnotationH>;<AnnotationW>
```

Where:

- `<Path>` is the path to the annotated image.
- `<OffsetX>` and `<OffsetY>` are the position of the upper left corner of the annotation.
- `<AnnotationH>` and `<AnnotationW>` are the height and width of the annotation within the image.

In both cases, the `path` is computed using a modified version of the open source Okfn annotator [7] range class. In order to improve cross browser compatibility, CAT replaces Okfn's XPath pointers with CSS selectors. There are two reasons for this change. Firstly, different browsers will render pages in different ways, which means that XPath is not always a reliable means of locating a specific element in the markup. Secondly, support for XPath has been removed from current releases of jQuery. CSS selectors, however, are still supported and hence are the more suitable choice.

Additionally, rather than using browser ranges, CAT uses Rangy [20] ranges. Rangy is an open source JavaScript library which creates a virtual representation of a selected range that is independent of the browser being used. Rangy can then map this virtual range to the current page, taking into consideration the browser being used. Pointers are generated with respect to this virtual range so that the result should always evaluate to the same document location regardless of the environment.

FAST provides a `pointer` field as part of an annotation's representation. This is a free-text field, allowing CAT to define its own format for indicating the section of a document with which an annotation is associated. The serialized representation of the annotated range is stored at this location.

4 Architecture

4.1 FAST Architecture

The FAST annotation service comprises three sub-systems:

- **logging infrastructure:** lays behind all the components of the FAST system, captures information such as the user name, the IP address of the connecting host,

the action that has been invoked by the user, the messages exchanged among the components of the system in order to carry out the requested action, any error condition, and so on. Moreover, as far as the FAST RESTful Web Application is concerned, it captures also the HTTP logs and represents them according to the W3C Extended Log File Format [15]. Furthermore, the log events can be accessed and searched interactively by means of (possibly) complex extended Boolean queries, comprising both exact and best match clauses, giving thus the possibility to mine and fully exploit them;

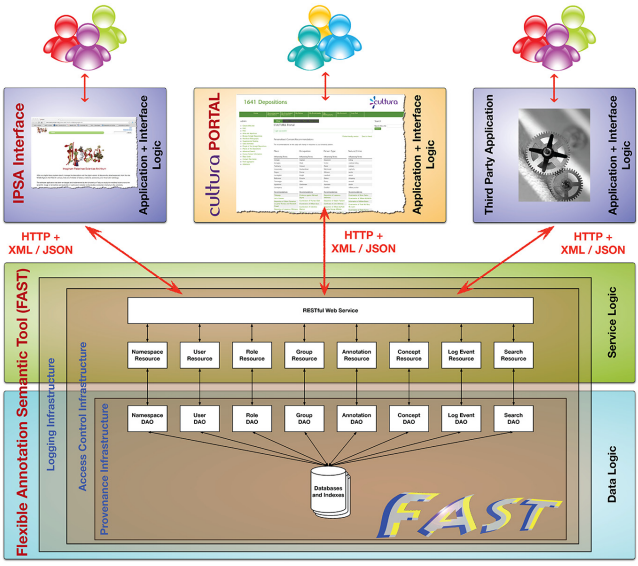


Fig. 3. Architecture of the FAST annotation service

- **access control infrastructure:** takes care of monitoring the access to the various resources and functionalities offered by the system. On the basis of the requested operation, it performs: (i) authentication, i.e. it asks for the user credentials before allowing to perform an operation; (ii) authorization, i.e. it verifies that the user currently logged in holds sufficient rights to perform the requested operation; The access control policies can be dynamically configured and changed over the time by defining roles, i.e. groups of users, entitled to perform given operations. This allows institutions to define and put in place their own rules in a flexible way according to their internal organization and working practices. Moreover, the access control infrastructure provides fine-grained control over the access to the specific resources, based on the permission granted to the resources, e.g. only the owner of a private resource and read it, even if the reading of that resource is granted to all roles;
- **provenance infrastructure:** keeps fine trace, for each resource managed by the system, of its full lineage since its first creation, allowing us to reconstruct its fully history and modifications over the time. Provenance events are statements about a

resource of the form: <when> <who> <predicate> <what> <why> where <when> is the time stamp at which the event occurred; <who> is the user who caused the event; <predicate> is the action carried out in the event, i.e. CREATED, READ, or DELETED; <what> is the resource originated by the event, i.e. a dump of the actual content of the resource; and <why> is the motivation that originated the event, i.e. the operation performed by the system that led to a modification of the resource. For all these events, a dump of a resource is stored in the Provenance Infrastructure, thus allowing us to access to the different versions of it over the time, even after it has been deleted from the system.

The FAST annotation service is exposed as a RESTful Web Service [16] which allows for the development of different applications and plug-ins over it in an open, collaborative, and scalable way which ensure sustainability over the time.

The FAST annotation service has been developed by using the Java[3] programming language, which ensures good portability of the system across different platforms. We used the PostgreSQL[4] DataBase Management System (DBMS) for the actual persistence of annotations and its full text extension for indexing and searching the full text components of the managed resources. The Apache Tomcat[5] Web container and the Restlet[6] framework have been used for developing the FAST RESTful Web Application.

4.2 CAT Architecture

The architecture of the CAT annotation tool is comprised of two layers; A client-side front end, coded using JavaScript and jQuery, and a Drupal 7 module back end, written in PHP.

The front end runs in the user's browser and provides them with a user interface through which they can interact with annotations. When a user has chosen a particular course of action, the data is passed into the logic module where their request can be processed. Depending on the nature of the request, certain third party libraries may be used in the procedure. For example, in the process of annotating a text object, the location of the text in the document must be recorded in a cross platform manner. In order to do this, a representation of the highlighted range is generated using rangy. This is a purely virtual range which means it is slightly slower than using the browser's range, but it has the advantage of being cross platform. Using a modified version of the Okfn path finder, the logic then computes a serialized path to the selected location represented by rangy which can be stored as a pointer in FAST. When annotating images, the process is the same except that jCrop [21] provides details of the selected region rather than Rangy. Retrieving an annotated region is simply the reverse of this process.

The representation of an annotation created here is a simplified version of the FAST description of the annotation. This is to minimize the amount of data that a user

[3] http://www.oracle.com/technetwork/java/index.html
[4] http://www.postgresql.org/
[5] http://tomcat.apache.org/
[6] http://www.restlet.org/

must send and receive to and from the server. For example, details such as namespaces are added on the back end rather than on the front end (and thus are managed by the site administrator). Furthermore, when managing details such as groups, the user's permissions are derived from the verbose annotation description on the server and then passed as a single value in the simplified representation.

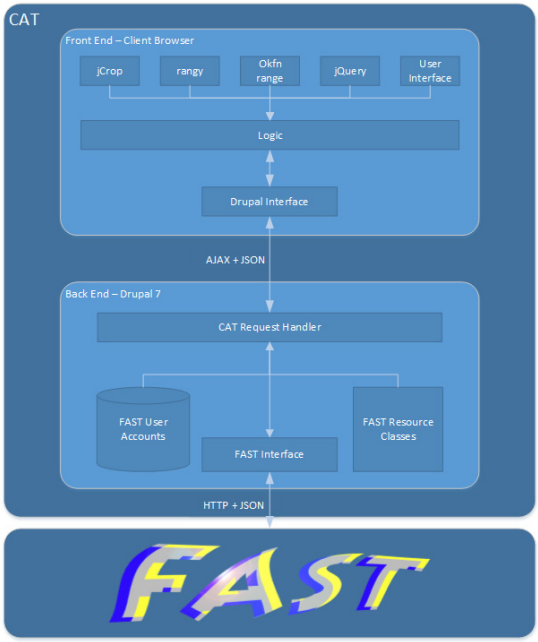


Fig. 4. Architecture of the CAT annotation tool

The Drupal 7 module on the back end acts as a relay between FAST and the user. Requests for annotation creation, deletion, download etc. are passed from the front end to a request handler function on the back end. This callback function structures the data sent by the front end so that it conforms to the FAST schema and then generates the HTTP packets to be transferred. There is some logic applied at this point to determine which packets need to be sent and in what order for the request to be fulfilled. Once the system is ready, the packets are sent on to FAST. The Drupal module then waits for a response from the remote service. When one is received, the result is returned to the front end via the same callback function through which the request was initially made.

The choice of a Drupal module as a means of implementation means that adding FAST-CAT to any site using the Drupal CMS should be a very simple process. Additionally, as the Drupal module is only acting as a relay, it should be a relatively simple process to swap out the back end for a more server agnostic implementation, allowing FAST-CAT to be deployed on any website, rather than only those using the Drupal 7 content management system.

Certain requests such as creating and viewing annotations require user authentication by FAST. As FAST is a stand-alone service, it maintains its own record of user accounts and login details. This means that for each user who is registered on the CULTURA site (see section 5), a separate account must be created for them in FAST. CAT performs this registration automatically.

5 The CULTURA Environment

CULTURA[7]is a three year, FP7 funded project, scheduled to finish in February 2014. Its main objective is to pioneer the development of personalised information retrieval and presentation, contextual adaptivity and social analysis in a digital humanities context. In its current form, it aims to provide adaptive and personalized access to two historical collections – the 1641 depositions [17] and IPSA [18].

FAST-CAT has been integrated into the environment in order to provide users with an additional means of interacting with the portal, as well as to provide some feedback for CULTURA's user model regarding a user's interests. At present, CULTURA (and by extension FAST-CAT) is being evaluated by three groups of users.

A team of MPhil students and professional researchers from Trinity College Dublin are using FAST-CAT as part of their teaching, collaboration and research into the 1641 depositions. These users will be testing the annotation tool in a free form manner. How they choose to annotate and what content they label is entirely determined by their own needs.

The 1641 depositions are a collection of handwritten witness statements taken from Protestant men and women of all classes of society during the Catholic rebellion of 1641. These documents provide an incredible insight into the state of Ireland, Scotland and England in the period surrounding the rebellion and are an unparalleled source of information in this field. The depositions are textual in content, so these students will serve only to evaluate the text annotation aspect of the tool.

Providing an alternative insight to FAST-CAT is a group of secondary school students from Lancaster who used the annotations as part of a project they were given during a lesson. Their experience was more guided than that of the masters students as they were directed to highlight information or points of interest using FAST-CAT and then deliver a presentation using annotations to help with organization. The focus of this lesson was on the 1641 depositions.

Masters students in Padua will test the image annotation functionality of FAST-CAT as part of their research into the Imaginum Patavinae Scientiae Archivum (IPSA) [18] collections of illuminated manuscripts.

The IPSA manuscripts are a series of illustrated documents which describe the various properties of herbs and plants dating from as far back as the 14th century. They have the very rare and wonderful quality of having been incredibly accurately and realistically hand drawn from nature. While there is a Latin commentary for each plant, the real interest in these documents lies in the illustrations.

[7] http://www.cultura-strep.eu/

Similarly to the MPhil students, the approach of these masters students to annotating documents will be determined by their own research methodology. The intention is not to guide the users on how to use FAST-CAT, but rather to make them aware of the functionality provided and observe how they choose to apply it.

The various features offered by FAST-CAT and its user interface will be evaluated in detail and comparisons will be drawn between the manner in which different user groups availed of annotations depending on their level of expertise and the type of documents examined. Furthermore, FAST-CAT will also help to drive CULTURA's comprehensive user model by providing the site with updates on the user's behaviour regarding document annotation.

6 Conclusions

It is the belief of the authors that FAST-CAT has huge potential as an annotation tool within the digital humanities field. However, it is still a young tool with much room for future expansion and enhancement. Some of the required additions are already known and are currently being developed. Others will be dependent on user feedback from test groups as they identify issues the experienced within their domains.

A large facet of plans to improve FAST-CAT is to increase the range of content types with which it may be used. At present, it provides for the annotation of text and images. Possible additions to this list include dynamic content types such as SVDs.

As was mentioned in section 4.2, it is possible to make FAST-CAT more server agnostic by swapping out the Drupal 7 back end for a more general php script. It is expected that this script will be developed and provided with future versions of FAST-CAT so as to increase the range of portals to which it may be applied

Further to this, another part of the future development of FAST-CAT will be focused on improving the user's experience. It is intended that the tool be as intuitive and easy to use as possible. How this will be achieved is to be this based on the feedback given by the user groups during the CULTURA trials.

Acknowledgements. The CULTURA (contract no. 269973) and the PROMISE[8] network of excellence (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

References

1. Agosti, M., Ferro, N., Frommholz, I., Thiel, U.: Annotations in Digital Libraries and Collaboratories – Facets, Models and Usage. In: Heery, R., Lyon, L. (eds.) ECDL 2004. LNCS, vol. 3232, pp. 244–255. Springer, Heidelberg (2004)
2. Agosti, M., Bonfiglio-Dosio, G., Ferro, N.: A Historical and Contemporary Study on Annotations to Derive Key Features for Systems Design. International Journal on Digital Libraries 8(1), 1–19 (2007)

[8] http://www.promise-noe.eu/

3. Agosti, M., Ferro, N.: A Formal Model of Annotations of Digital Content. ACM Transactions on Information Systems (TOIS) 26(1), 3:1–3:57 (2008)
4. Bélanger, M.-E.: Ideals (February 10, 2010), https://www.ideals.illinois.edu/bitstream/handle/2142/15035/belanger.pdf?sequence=2 (retrieved October 25, 2012)
5. Barbera, M., Meschini, F., Morbidoni, C., Tomasi, F.: Annotating digital libraries and electronic editions in a collaborative and semantic perspective. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 45–56. Springer, Heidelberg (2013)
6. Ferro, N., Silvello, G.: Empowering Archives through Annotations. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 57–68. Springer, Heidelberg (2013)
7. Okfn (n.d.). Okfn Annotator, http://okfnlabs.org/annotator/ (retrieved June 2012)
8. TILE, TILE: text-image linking environment (2011), http://mith.umd.edu/tile/ (retrieved July 2012)
9. Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., Wade, V.: The CULTURA Project: Supporting Next Generation Interaction with Digital Cultural Heritage Collections. In: Ioannides, M., Fritsch, D., Leissner, J., Davies, R., Remondino, F., Caffo, R. (eds.) EuroMed 2012. LNCS, vol. 7616, pp. 668–675. Springer, Heidelberg (2012)
10. Hampson, C., Lawless, S., Bailey, E., Yogev, S., Zwerdling, N., Carmel, D., Conlan, O., O'Connor, A., Wade, V.: CULTURA: A Metadata-Rich Environment to Support the Enhanced Interrogation of Cultural Collections. In: Dodero, J.M., Palomo-Duarte, M., Karampiperis, P. (eds.) MTSR 2012. CCIS, vol. 343, pp. 227–238. Springer, Heidelberg (2012)
11. Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H.: The DELOS Digital Library Reference Model. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy (2007), http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf
12. Salton, G., Fox, E.A., Wu, H.: Extended Boolean Information Retrieval. Communications of the ACM (CACM) 26(11), 1022–1036 (1983)
13. Ferro, N.: Annotation Search: The FAST Way. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 15–26. Springer, Heidelberg (2009)
14. OASIS Search Web Services Technical Committee, searchRetrieve: Part 5. CQL: The Contextual Query Language Version 1.0. (2012), http://docs.oasis-open.org/search-ws/searchRetrieve/v1.0/searchRetrieve-v1.0-part5-cql.pdf
15. Hallam-Baker, P.M., Behlendorf, B.: Extended Log File Format – W3C Working DraftWD-logfile-960323 (1996), http://www.w3.org/TR/WD-logfile.html
16. Fielding, R.T., Taylor, R.N.: Principled Design of the Modern Web Architecture. ACM Transactions on Internet Technology (TOIT) 2(2), 115–150 (2002)
17. Trinity College Dublin, 1641 Depositions, http://1641.tcd.ie/index.php (retrieved November 2012)
18. Universita Degli Studi Di Padova, IPSA – Imaginum Patavinae Scientiae Archivum, http://www.ipsa-project.org/ (retrieved November 2012)
19. Rangy, http://code.google.com/p/rangy/ (retrieved November 2012)
20. Deep Liquid, jCrop, http://deepliquid.com/content/Jcrop.html (retrieved November 2012)

EuropeanaLabs: An Infrastructure to Support the Development of Europeana

Nicola Aloia, Cesare Concordia, Carlo Meghini, and Luca Trupiano

Istituto di Scienza e Tecnologie dell'Informazione,
National Research Council, Pisa, Italy
{nicola.aloia,cesare.concordia,carlo.meghini,
luca.trupiano}@isti.cnr.it

Abstract. This document describes the Europeana Development and communication infrastructure, called EuropeanaLabs, built inside the EU projects Europeana and Europeana v. 2. The EuropeanaLabs consists of a number of servers, storages and communication devices; it is used to create Virtual Machines, called sandboxes, used by Europeana foundation communities. EuropeanaLabs provides a test environment, for applications and demos, to several cultural heritage and technology projects, most of them funded by EU, in addition it features a set of servers for cooperative work. In this paper we present the general architecture of the EuropeanaLabs infrastructure.

Keywords: Digital Library infrastructure, Europeana, sandbox.

1 Introduction

In the wide public, the Europeana Digital Library is primarily perceived as a portal exposing a great amount of cultural heritage information. Even though this perception is not entirely misleading, Europeana is much more. More precisely Europeana is an open services platform enabling users and cultural institutions to provide, manage and access a very large collection of information objects representing digital and digitized content [9]. Europeana is also a set of resources and tools open to the Community for developing, creating and disseminating new information resources. The set of these instruments provides the infrastructure, called EuropeanaLabs, on which is based the development of Europeana. The EuropeanaLabs provides development environments and various tools for building and managing digital libraries. These tools range from application servers for harvesting data and metadata (eg Repox), to Customers Relationship Manager (CRM), to collaborative work environments, products management, software validation, application showcasing, etc. From the organizational point of view the Europeana DL is the result of a number of activities run by different actors located across Europe, coordinated by the Europeana Digital Library Foundation (EDLF).

To define such a complex organization we can use the classical definition of Information System (IS): “a combination of Information Technology (IT) and people's activities that support operations, management and decision making [8]”.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 53–58, 2014.

According to this definition, the Europeana Portal is a component of the system, more specifically it is a web application using the Europeana API to provide services, in particular content discovery, to the Europeana Digital Library.

This paper focuses on, the computer system that act as a backend for EuropeanaLabs.

2 The EuropeanaLabs

As previously mentioned the Europeana DL is a collaborative work coordinated by EDLF and, up to today, the most part of the activities are carried on inside specific EU funded projects.

From the organization point of view this means that the community of users working on Europeana is composed by autonomous teams, usually geographically distributed, and number and needs of the community do vary over time.

We can individuated the following main features of the community of users working on Europeana:

- Distribution: there are geographically distributed teams, working on separate processes, often needing a close interaction each other to execute their activities.
- Heterogeneity: different teams can use different approaches for the same activity. For instance the various development teams could use different development methodologies.
- Scalability: at any moment new teams can join (or retire from) the organization
- Autonomy: every team must have a complete, autonomous working environment.

Up to November 2012 the community working on Europeana is composed by 1660 registered users, grouped in 42 main teams and each team works on one or more Europeana related activity.

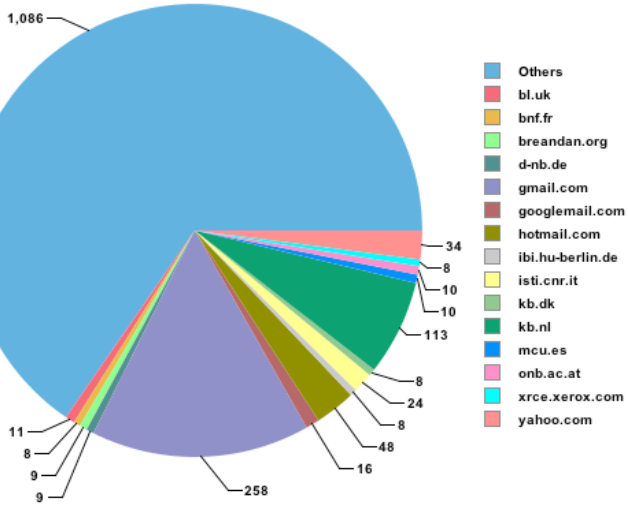


Fig. 1. The distribution of EuropeanaLabs community by email domains

In such a scenario, the role of EDLF is crucial. It is EDLF that coordinates activities of the Europeana community, collects and validates results, publishes them in Europeana (in form of new content or as new Europeana services).

The EuropeanaLabs has been designed to implement all needed tools and facilities to enable community users and EDLF to carry on their tasks.

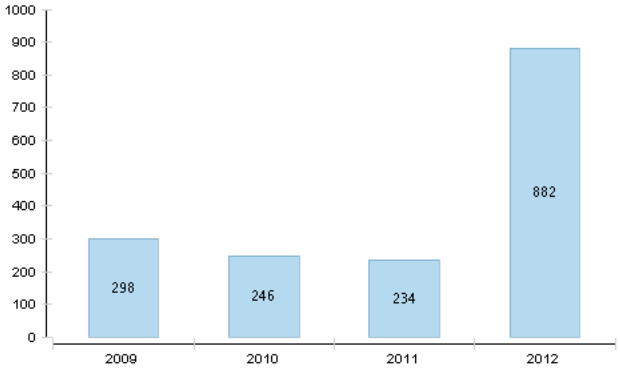


Fig. 2. New users registered in EuropeanaLabs every year

2.1 The Europeana Sandboxes

Basically the EuropeanaLabs provides computational and storage resources in form of autonomous systems called Europeana sandboxes.

Generally speaking, the term sandbox, in computing, may refer to:

- an isolated runtime environment used to run untrusted code (security sandbox)
- a testing environment that isolates untested code changes and outright experimentation from the production environment or repository, in the context of software development including Web development and revision control (development sandbox)[1].

We designed Europeana sandboxes by joining features of both categories; essentially a Europeana sandbox has the following features:

- it is a complete host computer on which a conventional operating system boots and runs
- it provides a controlled set of resources (main memory, storage, cpu, etc) and is accessible via the network
- it can be easily migrated between different hardware servers.

Every team working on one Europeana project, can ask for one or more sandboxes for tasks related to their activity; if the request is accepted, sandboxes are created and assigned to the team.

Europeana sandboxes are implemented using Hardware virtualization [6]: every sandbox is a Virtual Machine (VM). This is an obvious choice, virtualization paradigm

enable us to fully implement the specification defined by the project. The challenge is to define an architecture reliable, scalable and flexible enough to support the Europeana ambitious goals.

2.2 Architectural Description

The EuropeanaLabs infrastructure has been designed, developed and is maintained by CNR-ISTI. The work started in 2008: the initial configuration comprises a small server and a backup storage.

Over the years the complexity of the infrastructure has increased with the complexity of the project. It has finally conformed to a highly flexible, scalable and robust model. Among typical requirements for complex information systems (reliability, fault tolerance, etc) scalability is a major one for EuropeanaLabs: the infrastructure must be easily upgraded or remoulded to be adapted to the changing demands. The actual EuropeanaLabs model can really fulfil the current resources needs and easily meet the future ones, without requiring substantial changes. It comprises four main logical components:

1. the computational component,
2. the storage component,
3. the network component,
4. the administration component.

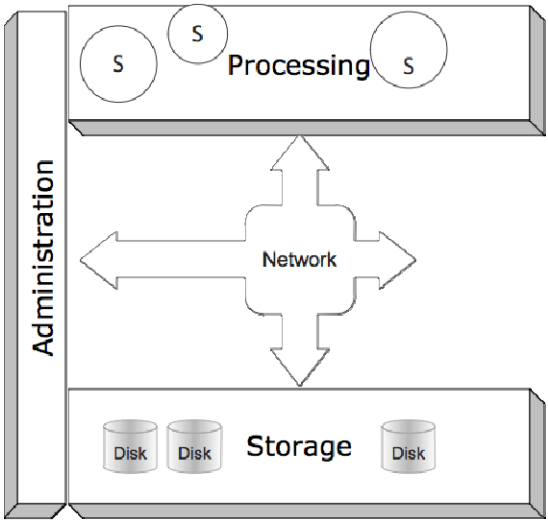


Fig. 3. EuropeanaLabs infrastructure components

The **computational component**, composed by 4 virtualization servers, is responsible of creating and running all the sandboxes and provides them with CPU and main memory. It also handles the storage space, remotely provided by the storage

component, making it available to the sandboxes. Finally, it provides Internet access to the sandboxes.

The **network component**, is composed by 2 high speed network switches. It provides LAN services and Internet connectivity to both the servers and the sandboxes. The data traffic between the virtualization servers and storage servers is done through standard protocols (AOE and iSCSI) and isolated from the Internet traffic, thus forming a storage area network (SAN). All network connections are made via a redundant structure both at the level of apparatuses that of physical links to provide fault tolerance and modularity.

The **storage resources**, composed by 2 servers, are redundant (all disks are in RAID configurations) to provide fault tolerance. They have grown with the project both in size and technology, and moved from general-purpose operating systems to dedicated storage systems with specialized hardware with the aim of improving performance and modularity.

The management and monitoring facilities, which make up the **administration component**, are provided by various software applications hosted on dedicated workstations. The management part makes use of standard hypervisor's tools, customized scripts and multiple servers management tools. The monitoring part uses Munin [3] to actively check all the relevant resources and services using a software agent installed on every virtualization and storage server. Another tool, Xymon [4], is used to check the status of the relevant network services of each sandbox without installing any software agent. The Internet traffic from and to the sandboxes is monitored in real time using tools like nProbe and nTop [5].

As all the sandboxes can be run on every virtualization server and have disk space provided by every storage server, this kind of architecture is reliable, flexible and easily expandable: virtualization servers can be added to increase the number of sandboxes, storage servers to increase the disks size, network switches to increase the number of connection links. If needed, all the sandboxes can be moved between virtualization servers, and the resources allocated to them can be changed.

2.3 EuropeanaLabs Software

The software used in the EuropeanaLabs infrastructure is open source. Besides from the servers operating system (Debian and OpenIndiana), the infrastructure needs hypervisoring, managing and monitoring software. “Hypervisor or virtual machine manager (VMM) is a piece of computer software, firmware or hardware that creates and runs virtual machines” [7]. As virtualization hypervisor Xen [2], being a widely tested and robust software, was chosen and installed on every virtualization server.

The grown of the infrastructure over time and the analysis of its usage by the community has resulted in the need for a dedicated middleware to manage and monitor all its resources, which can make the infrastructure still more flexible to adapt to the users requirements. We are currently working on this.

3 Conclusions and Future Works

The infrastructure, after five years of deployment, has been set up to fulfil all Europeana requirements. It currently provides enough computational, storage and communication resources to meet the needs of the project. It has the flexibility that allowed different reconfiguration to be done when needs have changed, has proved to be reliable with no data loss reported, has experienced minimal downtime without impacting the work of the Europeana community participants.

However some work still has to be done in order to make the infrastructure more flexible and manageable, in particular we're currently implementing the following features:

- live migration of sandboxes from any server to any other,
- monitoring and management operations done by a centralized middleware which would be able to define, create, migrate and destroy virtual machines and dynamically reallocate all the physical resources.

Another mayor activity being carried on by ISTI team is the design and implementation of a new and dedicated user interface for monitoring the infrastructure's resources status and usage, aggregating existing control data in a different way and paying particular attention to dependencies between sandboxes.

References

[1] http://en.wikipedia.org/wiki/Sandbox_(software_development)
[2] Xen, http://www.xen.org/
[3] Munin, http://munin-monitoring.org/
[4] Xymon, http://xymon.sourceforge.net/
[5] nTop, nprobe, http://www.ntop.org/
[6] http://en.wikipedia.org/wiki/Hardware_virtualization
[7] http://en.wikipedia.org/wiki/Hypervisor
[8] Definition of Application Landscape. Software Engineering for Business Information Systems (sebis) (January 21, 2009)
[9] Concordia, C., Grad-mann, S., Siebinga, S.: Not just another portal, not just another digital library: A portrait of Europeana as an application program interface. IFLA Journal 36(1), 61–69 (2010)

A Digital Infrastructure for Trustworthiness

The Sapienza Digital Library Experience

Angela Di Iorio[1], Marco Schaerf[1], Maria Guercio[1], Silvia Ortolani[1], and Matteo Bertazzo[2]

[1] Sapienza Università di Roma, Rome, Italy
{angela.diiorio,marco.schaerf,maria.guercio,
silvia.ortolani}@uniroma1.it
[2] CINECA, Bologna, Italy
m.bertazzo@Cineca.it

Abstract. The building process of Sapienza Digital Library's (SDL) digital resources was designed for collecting the information required by the Open Archival Information System (OAIS) Preservation Description Information(PDI): Provenance, Reference, Fixity, Context, and Access Rights Information. The Submission Information Packages'(SIP) preservation metadata was encoded in the semantics of the PREMIS standard which is the implementation metadata set, mapped from the OAIS conceptual model. The conformant implementation of the PREMIS standard was one of the principles which permeates the SIP building process. All relevant legal aspects and formal agreements, referred to the organizations involved in the different OAIS functions of the SDL digital repository, were analyzed and structured for their inclusion into the forthcoming AIP management, and for unleashing of the preservation strategies, and for supporting the authenticity of resources.

Keywords: DL Architectures and infrastructures, Long term preservation, Metadata creation, management, and curation, OAIS, METS, PREMIS.

1 Introduction

The Sapienza Digital Library[1] (SDL) is a research project undertaken by Sapienza Università di Roma (Sapienza), the largest Europe's campus, and the Italian supercomputer center Cineca[2], the 9th in the Top500[3], which is a no profit consortium, made up of 54 Universities, 2 Research Institutions and Ministry of Education, University and Research.

The SDL project aims to build an infrastructure supporting preservation, management and dissemination of the past, present and future digital resources, containing the overall intellectual production of the Sapienza University[3].

[1] Sapienza Digital Library http://sapienzadigitallibrary.uniroma1.it (expected on May 2013)
[2] Cineca consortium http://www.cineca.it
[3] Top500 supercomputer sites http://www.top500.org/

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 59–69, 2014.

The digital infrastructure set up by the SDL project was build aiming at the conformance to the OAIS(ISO 14721:2003) functional model[1] and developing compliant services supporting the Long Term Digital Preservation (LTDP).

Both two projects participants have built two repositories that can be defined in OAIS terms like "Cooperating: Archives with potential common producers, common submission standards, and common dissemination standards, but no common finding aids". The interchanging repositories share a common metadata infrastructure based on the most spread metadata standards for digital libraries, and the Information Packages are replicated in both repositories.

The provision of a SIP, equipped with the PDI required by OAIS, was considered an essential requirement in the design of both digital repositories and in the design of the metadata framework on which is based the IP exchange.

2 The SDL Preservation Technical Strategy

Sapienza's organizations, or other organizations in legal agreement with a Sapienza's organizations, will provide digital content for the Sapienza Digital library services supporting the digital curation activities. The SDL project agreement between Sapienza and CINECA has established the commitment of both in making up the services for digital preservation. Regarding to the preservation services, the replication of storage, geographically dispersed, is one of the technical strategy for the trustworthiness[4] of the overall system.

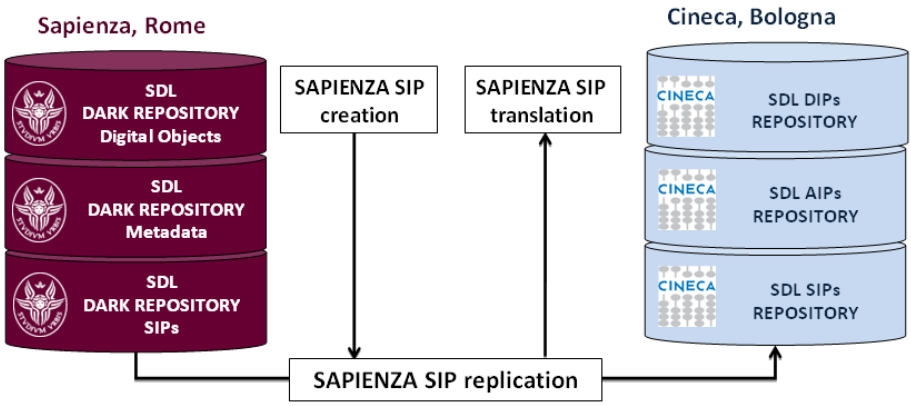


Fig. 1. SDL cooperating repositories, geographically dispersed

The first repository, where Sapienza SIPs are created, is located in Rome at the Sapienza University. Every Sapienza SIP created is then replicated into the CINECA storage, which is located in Bologna.

The Sapienza SIP replicated is ingested and translated into corresponding AIP and DIP, that are managed by the SDL management system based on Fedora Commons[4]. Both systems (Sapienza and Cineca) share semantics about the common standardized description of the original SIP, produced by the Sapienza University.

The Sapienza SIP contains metadata tailored on metadata documents' models, that CINECA technological system translates in provision of services for archiving and dissemination.

The technical level of interaction of the SDL and CINECA archives can be defined, in OAIS terms, as *cooperating* archives considering that the performed activities are based on a standard agreement and they have common SIP and DIP format and related communities of interest. The Sapienza SIPs produced by the University and stored in the Sapieza local dark archive, is replicated in the CINECA archive and ingested and translated in the SDL Archival Information Package and the corresponding DIP, updated with the Events information and provided on request.

As the OAIS specifies "The only requirement for [the Cooperating Archives] architecture is that the cooperating groups support at least one common SIP and DIP format for inter-Archive requests", the SDL framework was designed on metadata specifications that are commonly used for SIP and DIP in the Sapienza-Cineca interchange scenario. In order to support the standard agreement cooperation, "a set of mutual Submission Agreements, Event Based Orders, and user interface standards to allow DIPs from one Archive to be ingested as SIPs by another"[1] was designed, and at this moment is under implementation.

3 Designing the SDL metadata framework for LTDP

The metadata framework conceived for SDL has respected the following requirements oriented to support the LTDP:

- conformant with OAIS, in order to support the OAIS model of information, to fulfill the responsibilities for operating an OAIS Archive, and to underscore the trustworthiness of holding repositories[1];
- prearranged to hold different standard descriptions on which implementing future integration services, supporting the use of wide-ranging knowledge's materials for different designated communities;
- prearranged to the exchange with other digital library systems or other information management systems, maintaining the information about provenance;
- prearranged to the LTDP and equipped with the essential metadata, enabling the long term management.

The arrangement of consistent information supporting the LTDP has followed the structure of the Preservation Description Information (PDI), which is composed by information about PROVENANCE, REFERENCE, CONTEXT, FIXITY, and ACCESS RIGHTS.

[4] Fedora Commons Repository System `http://fedora-commons.org/`

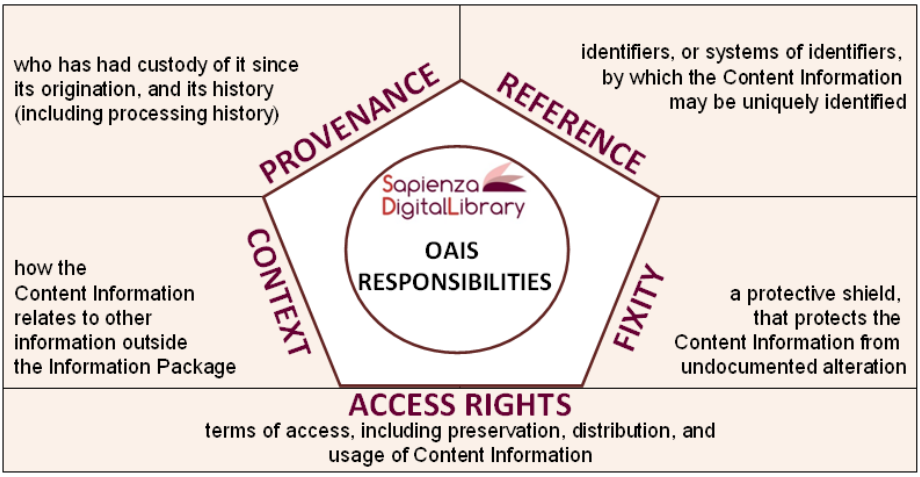


Fig. 2. SDL responsibilities discharging based on a consistent PDI

3.1 The CONTEXT and PROVENANCE Information

The CONTEXT information contains pointers to its environment by means of structured information referred to the originating organization (Sapienza's organizations libraries, museums, investigations departments). The CONTEXT information documents "why Content Information (CI) was created and how it relates to other CI objects existing elsewhere"[1].

The PROVENANCE information, which describes the source of CI, and in particular, "who has had custody of it since its origination, and its history (including processing history)" is provided by the Sapienza's organizations (both domain specific and technical) that produce, own, manage or have the custody of the CI.

Because Sapienza's University is a public institution, usually the business rules, for holding intellectual material, follow national or legal rules, like for example the Italian Author's Rights (civil law[5]) for the Intellectual Property information, or the Italian National Librarian System cataloguing rules for describing CI. It means that Sapienza's organizations, as public bodies, do already provide information following rules publicly and legally established.

Furthermore the provision of consistent CONTEXT and PROVENANCE information makes feasible to sustain both evidence to support the Authenticity of the resources, and the Trustworthyness of repository.

The system's characteristic of providing authenticity evidence, based on the assurance about the reliability of the CI, strongly consists of the ability of acquiring and maintaining unambiguous information about the CONTEXT and PROVENANCE of the managed digital resources.

[5] http://it.wikipedia.org/wiki/Civil_law

3.2 The ACCESS RIGHTS Information

The ACCESS RIGHTS information and documentation corpora (access restrictions, legal framework, licensing terms, and access control) were gathered, selected, modeled, identified and referenced to their own CONTEXT and PROVENANCE information.

The ACCESS RIGHTS Information in SDL contains the access and distribution conditions stated in the Submission Agreement, related to the third party usage and the SDL management, distribution, dissemination and preservation. The Submission Agreement involves both organizations with project responsibility: Sapienza and Cineca.

It also includes the specifications about the application of rights enforcement measures:

- Identification of the properly authorized Designated Community (Access Control, e.g. the access to some SDL objects is allowed to the Sapienza's community, the submission of resources is allowed to specific Sapienza's communities…)
- Permission grants for preservation and for distribution and dissemination (Copyright information)
- Pointers to FIXITY, CONTEXT and PROVENANCE Information
- Information about digital inhibitors like signatures, passwords and other access control mechanisms applied at submission and preservation time
- Legal and licensing framework(s)

The different layers of terms of agreements and actions allowed in the different contexts of submission, preservation, management, access and distribution were properly identified and structured in the documentation system. All specific agreements signed by Sapienza's organizations, responsible for digital curation in SDL and the third party granting the digital content, are unambiguously identified by the system, stored and referenced by related digital resources and collections. The specific agreements are referred to the general agreement, involving Sapienza as owner institution of the Digital Library and Cineca as partner, providing specific technological services, which states the general terms of the standard agreement cooperation.

3.3 The REFERENCE Information

The REFERENCE information was based on an identification system conceived in consideration of the cooperative focus of the project. Every single object, resource, and collection must have the essential information for detecting the originating entity, and the custodian entity, which has the responsibility of the digital curation.

The identification system manages a mechanism for creating identifiers families that are strictly connected to the "real Sapienza's organization", which is responsible for the custody chain (digital curator). At every level of the digital resource, it is possible to get unambiguously information about the origin and, consequently, the history of the resource. Every single SDL object can be reused and repurposed in different contexts, and is provided of all bounding information about its PROVENANCE, and

its originating CONTEXT, which points to the PDI of the source, expressed at collection level. The opportunity of exploiting resources in a referable manner, also allows the flawless interchange with other repositories.

3.4 The FIXITY Information

Automatic production of FIXITYs information is provided at the early SIP creation stage. The FIXITY information is one of the technical requirement about the overall management of the digital resources accessioned by both SDL archives. This means that, likewise the REFERENCE information, at every layer of the SIP building the FIXITY information is automatically produced, following a bottom up method: from the single content objects, going up the metadata objects and finishing with IPs.

3.5 Discharging Responsabilities of the SDL Organizations

The design of the system at this moment was focused on the essential services of ingestion, archiving and dissemination, waiting the forthcoming implementation of a robust preservation management system. Nevertheless the SIP creation workflow was conceived for gathering all information necessary for supporting LTDP and covering information necessary to the "mandatory responsibilities that an organization must discharge in order to operate an OAIS Archive"[1].

The negotiation between SDL archive and the Sapienza's organizations is based on an agreement, which will formally cover all resources submitted to the digital repository. The agreement (at this moment in draft form) establishes the acquisition of properly selected CI, produced by Sapienza organizations, and requires the provision of the bare minimum of information necessary for a consistent PDI, specifically oriented to the Designated Communities.

A similar agreement model was defined for terms of services between Sapienza and external organizations, not belonging to Sapienza University. Those organizations, that are willing to donate resources and to use SDL services, need to sign a legal agreement with one of the Sapienza's organizations, which is declared as "digital curator" of the donated resources.

The aim of the agreement model is obtaining "sufficient control for preservation"[1], gathering copyright implications, intellectual property and other legal restrictions on use, and acquiring the right level of authority to modify Representation Information, in the future contexts of migration.

The Organizations responsible for the content are deputed to define its own Designated Community of consumers, with the support of the SDL Scientific committee, taking into account the harmonization of the domain specific information, with the existing SDL descriptive information, necessary "to enable the Designated Community to discover and identify material of interest" [7].

At this moment, the documentation about policies and procedures is not yet completed and is under revision, but will be easily integrated by the system once the responsible Organizations will be agreed on it. The SDL archive agreement will be also

integrated by the constraint which claims the conformance to the policies and technical rules established by the SDL management.

4 The Preservation Metadata Implementation

Considering the LTDP strategy adopted, the overall SDL SIP building workflow must ensure the basic provision of the preservation metadata, considered mandatory by the PREMIS standard[2], which is the preservation metadata framework mapped from the conceptual structure of the OAIS model. The SDL metadata framework was designed to guarantee the conformance with the PREMIS standard, both on semantic unit and data dictionary level. As stated by the conformance guidelines on the PREMIS implementation[5], the SDL framework design has followed requirements and constraints, defined in the PREMIS Data Dictionary[6], and the SIP building workflow has collected all metadata necessary to support the PREMIS conformance requirement(4.1).

The PREMIS Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process". The SDL PREMIS implementation was based on the underlying belief that, the trustworthiness of a repository system relies on the ability of tracking back information about the custodianship of the objects. The custodianship's history allows to trace responsibility's chains back to the agents responsible for the events that occurred in the digital history of the resources.

4.1 PREMIS Conformance Requirement

The PREMIS conformance declared by a repository system, means that the implementation of PREMIS information adheres to the principles of use, stated by the PREMIS conformance: use of semantic units, and use of the data dictionary.

The metadata elements held by Sapienza black repository management system, shares names and definition, and respects the use requirements of the PREMIS semantic units. The obligation, repeatability and application rules are respected (semantic units principle of use).

Moreover, all mandatory semantic units, related to the PREMIS entities, are supported and used by the repository system (data dictionary principle of use).

Consequently, the PREMIS internal conformance is respected and, at this stage of the implementation, the external conformance is supported just in the form of export. Further implementation will be allowing the import form.

The actual resources' metadata, archived in the SDL archives, are encoded and collected into the METS[6] container. The encoding of PREMIS semantic units is under development but the conformance level internal and external was already reached. All the information, deemed essential for supporting the trustworthiness of the system and the authencity of resources, are owned and managed by the Sapienza black repository

[6] www.loc.gov/standards/mets

management system, and the undergoing implementation will be encapsulating PREMIS semantic units into the existing metadata framework.

4.2 PREMIS Enrichment Workflow

The PREMIS enrichment workflow consists of those activities necessary to the information adjustement for extending the actual PREMIS internal conformance of the SDL system, toward the PREMIS external conformance. The workflow essentially shapes the SDL existing metadata in the form of export for cross-repository interactions.

The workflow essentially consists of the following activities, that gathers information, and enriches the base of data with the information needed:

- Detecting Intellectual Entities and assignment of the SDL identifier, created by the pertaining collection's identifier and a unique identifiers for the corresponding resource:
  - Identifiers coming from the originating records (bibliographic catalog or original database, spreadsheet…)
  - SDL record identifier assigned by the SDL resources acquisition function.
- Getting the information about Objects related to the Intellectual Entities. At this moment the information automatically gathered and provided by the system are more than that required by the PREMIS conformance:
  - a unique identifier for the object (type and value),
  - fixity information message digest, algorithm and the application used,
  - size,
  - format,
  - original name of the object,
  - information about its creation,
  - where and on what medium is stored,
  - relationships with other objects and other entities (via identifiers).
- Getting the information about the Events occurred in the lifecycle of the Objects until the SIP production:
  - a unique identifier for the event (type and value),
  - type of event (creation, replication, message digest calculation, validation),
  - date and time,
  - detailed description of the event,
  - a coded outcome of the event,
  - detailed description of the outcome,
  - agents (via identifiers), involved in the event and their roles,
  - objects (via identifiers), involved in the event and their roles.
- Getting the information about Agents, engaged in activities impacting on the Objects' digital history
  - a unique identifier for the agent (type and value),
  - agent's name,

- designation of the type of agent (person, organization, software),
- extended description of the agents connected to the Sapienza's organization context,
- events (via identifiers) that the agents has determined,
- rights statements (via identifiers), to which the agent is related.

- Getting the information about Rights statements that impact on the Objects management:
  - a unique identifier for the rights statement (type and value),
  - basis of right (copyright, license, statute, or other),
  - more detailed information about the rights statements,
  - actions allowed by the rights statement,
  - restrictions on the action(s),
  - term of grant, or time period in which the statement applies,
  - objects (via identifiers), to which the statement applies,
  - agents (via identifiers), involved in the rights statement and their roles.

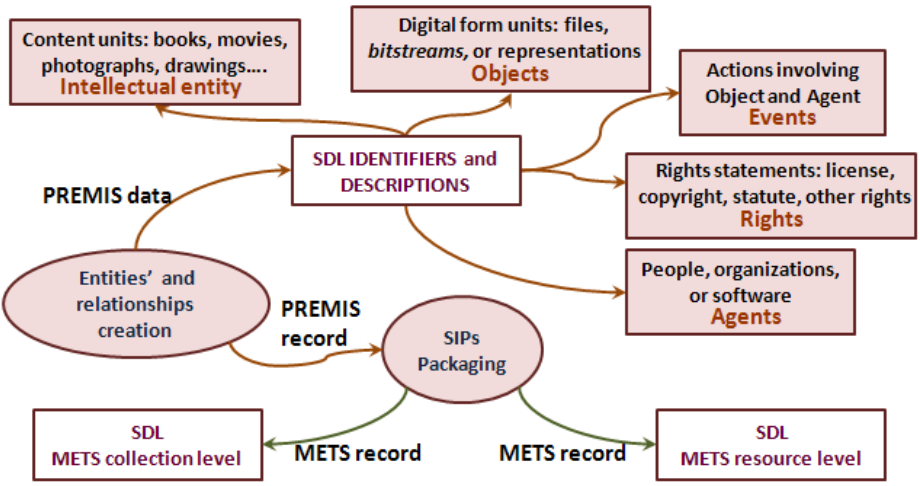


Fig. 3. Data flow diagram of the PREMIS entities implementation in the SDL metadata framework

In other words, if we express the metadata set information in natural language, it should result as: the Intellectual Entities, manifested[4] by different kinds of digital Objects, are produced by SDL Organizations made of people using tools. People, Organizations and tools are considered Agents responsible for specific actions. Actions are considered as Events in digital curation workflow, performed under specific conditions, formally defined and linked to the relevant Rights statements.

5 The Development of the Repository's Trustworthiness Value

Does the preservation metadata implementation support the trustworthiness of a system?

The maintenance of information conveying the digital history of the digital objects by means of preservation metadata related to OAIS[1], does support the future implementation of a trustworthy repository. The SDL SIP provision of a comprehensive set of preservation metadata, based on an international consensus-based standard like PREMIS, assure the availability of essential data for creating evidence of the trustworthiness of the archival systems. The AIP management, which will depends on the AIP data derived from the SDL SIP, will have a consistent set of information, available for implementing preservation services.

Any process of assessment, audit or certification strongly relies on the availability of consistent structured metadata. "Constant monitoring, planning, and maintenance of the repository, as well as conscious actions and strategy implementation will be required of repositories to carry out their mission of digital preservation" [7]. The cited management functions are strongly based on digital objects' metadata, that could negatively impact on the digital objects management, in case of absence, inconsistency, incompleteness.

Moreover, if the preservation metadata are encoded in PREMIS standard that is consensus-based of an international community of experts, the evidence of the repository's trustworthiness can be conveyed, by means of the consistent base of global standardized semantics, and on the hopeful alignment of the PREMIS conformance to the OAIS conformance.

In conclusion, the more the metadata framework will be updated, maintained and connected to the parallel "real" business evolution of the responsible Organization, showing evidence of the custody chain, the more it will be possible to have the result of the trustworthiness expected by the preservation digital repository.

References

1. Consultative Committee for Space Data Systems Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book (June 2012), http://public.ccsds.org/publications/archive/652x0m1.pdf
2. PREservation Metadata Implementation Strategies (PREMIS), http://www.loc.gov/standards/premis/
3. Di Iorio, A., Schaerf, M., Bertazzo, M.: Establishing a Digital Library in Wide-Ranging University's Context. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 172–183. Springer, Heidelberg (2013), http://link.springer.com/content/pdf/10.1007%2F978-3-642-35834-0_18

4. McDonald, R.H., Walters, T.O.: Restoring Trust Relationships within the Framework of Collaborative Digital Preservation Federations. Journal of Digital Information (2010), http://journals.tdl.org/jodi/index.php/jodi/article/view/757/645
5. PREMIS Editorial Committee, Conformant Implementation of the PREMIS Data Dictionary (October 2010), http://www.loc.gov/standards/premis/premis-conformance-oct2010.pdf
6. PREMIS Editorial Committee: PREMIS Data Dictionary for Preservation Metadata version 2.2 (July 2012), http://www.loc.gov/standards/premis/v2/premis-2-2.pdf
7. Audit and Certification of Trustworthy Digital Repositories. Magenta Book. Issue 1 (September 2011), http://public.ccsds.org/publications/archive/652x0m1.pdf

Historical Digital Archive and Geo-referenced Contents of the *Francigena Librari* Web Portal

Adriana Martinoli and Alfredo Esposito

General Direction for Libraries, Cultural Institutes and Copyright
Ministry for Cultural Heritage and Activities
{adriana.martinoli,alfredo.esposito-01}@beniculturali.it

Abstract. The General Direction for Libraries, Cultural Institutes and Copyright is promoting a project on the "Historical, cultural and religious itineraries valorization" and is achieving an institutional Portal dedicated to Italian Francigena track to create an unified gateway access for cultural and touristic contents concerning the Francigena route. The aim of the Portal is to offer news, events, cultural itineraries, interactive maps, virtual exhibitions, information, documents and images that represent the historical, cultural and religious heritage along one of the most important European pilgrimage routes, catalogued with innovative tools for managing taxonomies which allow advanced searches and geographical access to wide information. User may organize, in "Create your itinerary", his virtual or real journey sharing contents on the Francigena in the "community".

Keywords: Francigena route, Digitization, Libraries, Metadata, Indexing, GPS tracks, valorization, historical itinerary, cultural itinerary, religious itinerary.

The General Direction for Libraries, Cultural Institutes and Copyright is promoting a project on the "***Historical, cultural and religious itineraries valorization***" and is achieving an institutional Portal which is a unified gateway access for cultural and touristic contents[1,2,3]. The initiative starts with the Francigena Route project which aims to promote the itinerary described by Sigerico, Archbishop of Canterbury, in a travel diary which dates back to 990, containing the 79 stages back from Rome, where he received the *pallium* for the investiture by Pope Johannes 15th[1]. This project is part of an articulated institutional design whose goal is to preserve, to promote, and to make available the historical, cultural and religious heritage born, from age to age, on one of the most important pilgrimage routes. In order to provide broad visibility to

[1] President of the Ministers Council's Decree: Consults for Historical, Cultural, Religeous Itineraries, and institution of the scientific Committee of the Consults, 27th September 2007.

[2] Minutes of the Advisory Committee, 31st March 2009. In: www.francigena.beniculturali.it/web/valit/il-progetto

[3] Agreements significances of the legislative decree 22th january 2004, n.42, article 112 subscript between DGBID and the European Association of Francigena routes, the 18th December 2008, the 10th February 2010.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 70–76, 2014.

documents, initiatives, projects and resources committed along this common touristic-cultural and religious route, the portal has the following objectives:

- to spread the mapping of the cultural heritage through an articulated trace which connected Europe to the capital of Christianity;
- to involve regions, local government institutions, associations operating in the territory, ecclesiastical institutions, Council of Europe, including matters relating to cultural and touristic places of interest (accommodation, food and wine traditions, holidays and religious festivals, folklore, popular events, etc);
- to identify and share guidelines and operational standards;
- to make accessible cultural and tourist contents in digital format[4];
- provide visibility to the databases developed by subjects (individuals or corporations) involved in the project or who wish to join, creating a shared information network[5,6].

1 Digital Historical Archive

It represents the core of the Portal which includes data of documents and books preserved in libraries, archives and cultural Institutes. Data are organized in a logical and user-friendly way.

The browse page displays the following available searches: free search, advanced search, thematic areas, type of item, Institutions.

The advanced search allows refining by several approaches: a) names that includes people, institutions, congress, title. It appears the number of digital documents associated and then it is possible to narrow the set of results; b) topics; c) Place; d) Time.

The project implementation process involved different research and work stages for the selection of documents:

- Data analysis, collection and selection (bibliographic, archival, audiovisual and iconographic materials) equipped with standard descriptions [2-4], ;
- Description of the document/object aimed to the creation of Metadata necessary to the digitization process and to the web search of integrated information with other areas of the portal as well;
- Creation of rules for the automatic caption generation;
- Editing of captions and abstracts;
- Management of controlled and structured vocabulary for thematic indexing [5].

---

[4] 20th December 2005 n.120 Paper draft Directive MiBAC, Departement for research, innovation and organization, General Direction for technological innovation and promotion. Guidelines for co-ordinated communication plan of the web-site of the MiBAC Institutes about their accessibility and quality.

[5] Pontifical Council for Culture, `www.vatican.va/roman_curia/pontifical_councils/cultr/index.htm`

[6] Ministry Draft/Res(2010)53. Partial Enlarged Agreement (EPA) concerning the cultural Itineraries.

The Portal hosts a selection of cultural and touristic contents chosen in collaboration with the following libraries which the digitization of historical material has covered 40,140 images:

- Biblioteca Angelica, Roma – Images: 16479;
- Biblioteca Casanatense, Roma – Images: 988;
- Biblioteca d'Archeologia e Storia dell'Arte, Roma – Images: 5192;
- Biblioteca di Storia Moderna e Contemporanea, Roma – Images: 31;
- Biblioteca Nazionale Centrale di Roma – Images: 1638;
- Biblioteca Vallicelliana, Roma - Images: 15812.

It also includes materials preserved in: Istituto centrale per i Beni Sonori ed Audiovisivi, Roma (16); Biblioteca Nazionale Marciana, Venezia (198); Biblioteca della Società Geografica Italiana, Roma.

The back-office management of digital documents includes several steps: managing the editorial processing of digital content by assigning index entries referencing to the controlled vocabulary; Multilingual generating captions for viewing summary of digital documents; Referencing digital contents through geographical approach; Screen access to the controlled vocabulary; Browsing the controlled vocabulary management Tree to add terms and items.

The front-end page result of the digital document shows content information acquired [6-7-8-9-10].

2 Itinerary and Walkway

This section includes data collection and mapping development of the pedestrian path along the Francigena Route through the use of the download official guide that shows all the information about road book of the route, maps and GPS tracks.

The official walkway was born through these procedures:

- Collection of data mapping and development of the pedestrian path along the Francigena Route [11].
- Implementation of the official guide download that lists all the path information, maps and GPS tracks.
- Conclusion of the procedure for the validation of the Via Francigena official route with the letter signed by the Italian Minister of Cultural Heritage and Activities, Sandro Bondi and by the Minister for Agriculture, Food and Forestry Policies, Luca Zaia (11/11/2009).

Different or alternative paths are evaluated by a "Procedure for the assessment of changes to the pedestrian Francigena Way", whose objective is to define the mode of assessment and "certification" of the changes to the pedestrian path of the Francigena Way. Through the "Stages" you can find travel guides already divided into small steps of 20-25 km that the tourist can download (road book). There are also GPS tracks to download on their device in order to have the path that is already configured in your device.

The macro area “Along the Francigena Way” dedicated to cultural and tourist trails consists of:

- news, events, information, cultural tours and tourist exhibitions [12] editing and editorial content, united being multilingual and fully geo-referenced, queried and integrated
- the function "Create your itinerary" (definisci il tuo percorso) gives the user the possibility to define your own itinerary and then to create and print a practical guide containing all information about tourist and cultural attractions, digital documents, cultural events and cultural places (libraries, archives and Cultural Institutes).

The Portal shares web 2.0 functionality and potentialities. The Facebook profile is open to promote and communicate the initiative in this popular social network. Tagging functionalities is available to create your own path.

Moreover concerning the development of the web 2.0 we are confident to enlarged the functionalities considering the valuation and the propriety of the information to insert in the web-page portal [13].

3 Technological Framework

This section shows an overall view of the technological architecture and describes the used software, the RDBMS and the communication protocol employed between different modules of the Francigena system.

The technological platform uses open source software. The Via Francigena Portal, [14] was developed by computer company Engineering Ingegneria Informatica [15] and managed, in the last time, by computer company Inera s.r.l. [16].

The main component of the Portal Architecture is Liferay Enterprise Portal [17] that is employed as portlet container and as CMS of the whole system. Liferay works with a Search Engine SOLR for browsing and indexing the contents. It is also used to access the Digital Library and to make metadata and digital items available. The protocol is used to share Digital Library data through calls and responses web services (SOAP over HTTP).

Calls and responses of the web services (SOAP over HTTP) allows the sharing of Digital Library data.

Fig. 1 describes the Francigena system with its different subsets and their main interactions. This framework allows a great modular scalable through the use of standard protocols.

Digital Archives: provide a secure storage of information, metadata, items and digital contents. It is realized with the object-relational database management system (ORDBMS) PostgreSQL [18].

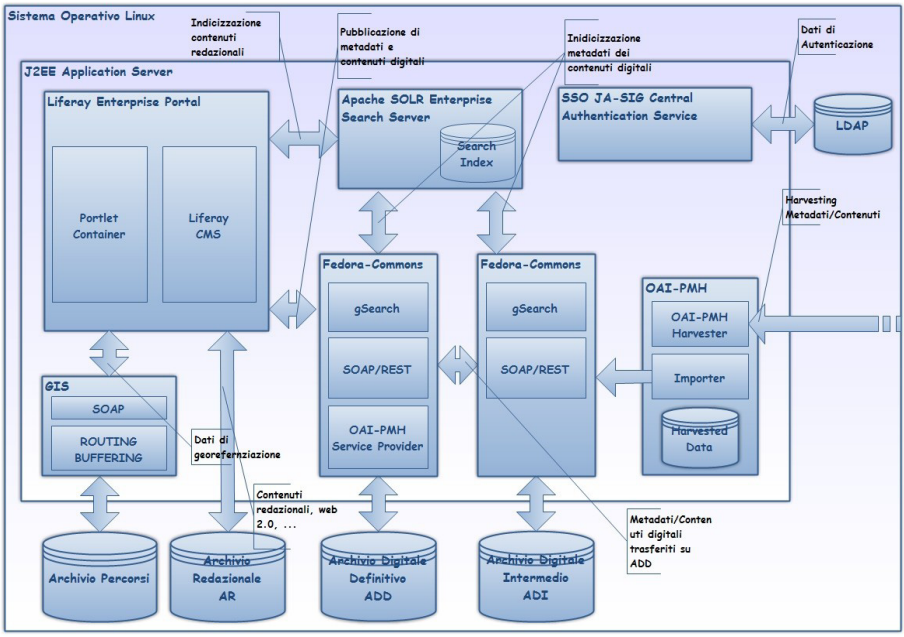


Fig. 1. The Francigena Portal architecture

OAI-PMH: is a mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH [2]. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP. The metadata and digital contents harvested are saved in the specific Digital Archive.

Enterprise Search Engine (SOLR) [19]**:** indexes all the metadata of digital content in the Archive Digital and the editorial content managed by the portal. The index then allows you to perform full-text searches on the various fields of metadata and editorial content in the system. The interaction is done through requests encoded in XML format over HTTP protocol.

Fedora-Commons (Digital Library - DL) [20]**:** it implements the functionality of the digital library which allow the administration and use of digital content and metadata. The description of the objects of the entire Fedora is based on the use of Administrative Metadata Management (MAG) to describe documents and reports and their relation. The Metadata are sent to the search engine for indexing when the medadata are inserting or editing or deleting. To do this it uses the form gSearch that translates and sends it to the indexing engine JMS messages generated by the management functions of the persistence of digital objects metadata. The DL exposes an OAI-PMH service provider that enables the system to the sharing of digital content and metadata with other repositories of digital content.

Geographic Information System (GIS): It is the system that manages and provides all data geo-referencing and cultural itineraries used by the Francigena Portal. It uses and manages the archive paths. The macro functionality exposed and used by the subsystem Portal (Liferay Enterprise Portal) are: routing that allows you to calculate the route between two or more geo-referenced points, buffering that is used to identify points at a certain distance from a given point and management (creation, modification, deletion) of geo-referenced points. The features are available through web services (SOAP over HTTP).

Single Sign on (SSO): it is a service that allows the propagation of authentication credentials to all integrated subsystems that require login.

LDAP: repository of the users of the system. It is powered by logging to the portal.

References

1. Via Francigena Portal, http://www.francigena.beniculturali.it
2. OAI-PMH, http://www.openarchives.org/OAI/openarchivesprotocol.html
3. Administrative Metadata Management (MAG), http://www.iccu.sbn.it/opencms/opencms/it/main/standard/metadati/pagina_267.html
4. Dublin Core Metadata, http://dublincore.org/
5. Thesaurus (index entries), http://thes.bncf.firenze.sbn.it/
6. Fallace, M.: Promotion of the historical, cultural and religious routes, and the help of information technology. In: Via Francigena, the Magazine of the Major Cultural Route of the Council of Europe, vol. 28, pp. 7–9 (2008)
7. Giannetto, M.: Minutes of the meeting of the scientific committee of the national Council of the historical, cultural and religious route, held on 31 March 2009. In: Via Francigena, the Magazine of the Major Cultural Route of the Council of Europe, vol. 29, pp. 6–13 (2009)
8. Bonito, S.: Agreement for the promotion of cultural routes (in conformity with legislative decree no. 42. art 112 of 22 January 2004). In: Via Francigena, the Magazine of the Major Cultural Route of the Council of Europe, vol. 31, pp. 17–26 (2010)
9. Bonito, S., Martinoli, A.: The Via Francigena in the framework of the European cultural routes: The web Portal of the Direction General for Libraries, Cultural Institute and Copyright. In: Via Francigena, the Magazine of the Major Cultural Route of the Council of Europe, vol. 33, pp. 18–21 (2011)
10. Esposito, A., Martinoli, A.: The multimedia historical archives of the "Francigena Librari" web Portal. In: Via Francigena, the Magazine of the Major Cultural Route of the Council of Europe, vol. 34, pp. 13–15 (2012)
11. BAICR consorzio cultura, http://www.baicr.it
12. Handbook on virtual exhibitions and virtual performances, http://www.indicate-project.org/getFile.php?id=412
13. Bonito, S.: The Via Francigena in the National Policy of the Italian Ministry for Cultural Heritage and Activities. In: 7th Final Workshop Per Viam Project – Pilgrim's Routes in Action Via Francigena and the Pilgrimage Ways. This is Europe, January 16-18 (2013)

14. Handbook for quality in cultural Web sites: improving quality for citizens. Project Minerva (2004)
15. Engineering Ingegneria Informatica, http://www.eng.it
16. Inera s.r.l., http://www.inera.it
17. Liferay Portal, http://www.liferay.com/home
18. PostgreSQL, http://www.postgresql.org/
19. Apache Solr, http://lucene.apache.org/solr/
20. Fedora Database, http://fedoraproject.org/it/get-fedora

Refering sites:
Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane – ICCU, www.iccu.sbn.it
Internet Culturale, www.internetculturale.it
World Digital Library, www.wdl.org
Europeana, www.europeana.eu

The Heritage of the People's Europe Project: An Aggregative Data Infrastructure for Cultural Heritage

Michele Artini, Claudio Atzori, Alessia Bardi, Sandro La Bruzzo, Paolo Manghi, Marko Mikulicic, and Franco Zoppi

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

Abstract. HOPE (Heritage of the People's Europe) is a "Best Practice Network" for archives, libraries, museums and institutions operating in the fields of social and union history. The project provides unified access to materials about the European social and labour history from the 18th to 21st centuries. HOPE proposes guidelines and tools for the management, aggregation, harmonisation, curation and provision of digital Cultural Heritage (CH) metadata and digital objects. Moreover, it offers to institutions joining the HOPE network an operational Aggregative Data Infrastructure (ADI) for the collection, aggregation and access of metadata records from CH content providers. The HOPE ADI is realized using and extending the D-NET Software Toolkit, an enabling framework for data infrastructures.

Keywords: cultural heritage, aggregation, metadata records, mapping, service-oriented architectures, data infrastructures, D-NET.

1 Requirements of the HOPE Community

The Heritage of the People's Europe project (HOPE) provides a unified entry point for the social and labour history from the 18th to the 21st century in Europe. It federates digital object collections from several major European institutions in the field. The community is willing to share an aggregated information space and deliver digital cultural objects, including videos (e.g. documentaries on labour movements), pictures (e.g. photos from Gulags), drawings (e.g. posters from the "Commune de Paris"), and archival documents (e.g. newspapers of migrants), in turn described by highly heterogeneous metadata representations. The goal is to group and interlink such objects in order to establish opportunities for a new cross-country, cross-institution social history background.

To this aim, the HOPE community requires an Aggregative Data Infrastructure (ADI) [6] able to handle a varying number of content providers, which in turn deliver several data sources, each dedicated to store metadata records and files relative to

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 77–80, 2014.

different object typologies. Indeed, as it often happens in the Cultural Heritage (CH) domain, content providers may deliver data sources whose objects belong to diverse sub-communities (in HOPE referred to as profiles), which in HOPE are: library, archive, visual, audio video. Although a profile marks a data source as including material of the same "semantic domain", distinct data sources may store objects of different formats (e.g. images, videos, audio, text material) and different descriptive data models and relative metadata formats. For example, librarians and archivists typically model their digital objects according to different data models and schemata (e.g. MARC[1] for libraries, and EAD[2] for archives), but each of them may have a variety of ways to describe their objects. Furthermore, data sources may export their content via several standard protocols, such as OAI-PMH, FTP, etc. At the end of the project, a total of about 900,000 metadata records will be aggregated, describing around 3,000,000 files in the CH domain. HOPE digital objects will be available from the IAHLI[3] portal and delivered to Europeana[4] as XML records in EDM format.

2 The HOPE Infrastructure

Institutions joining the HOPE network benefit of an advanced distributed ADI which enables them to enhance the quality and the visibility of the digital cultural objects they preserve. Moreover, the project also delivers a Shared Object Repository dealing with the management of digital files for HOPE partners who cannot afford the cost of a local object file store. It allows institutions to deposit their files and it automatically applies conversion algorithms to create files in standard formats and with sizes suitable for web dissemination.

The HOPE ADI. The HOPE ADI is implemented using the D-NET [3][4] Software Toolkit. D-NET is an open source, general-purpose software conceived to enable the realization and operation of ADIs and to facilitate their evolution in time. D-NET implements a service-oriented framework based on standards, where ADIs can be constructed in a LEGO-like approach, by selecting, customizing, and properly combining D-NET services. The resulting ADIs are systems that can be re-customized, extended (e.g. new services can be integrated), and scale (e.g. storage and index replicas can be maintained and deployed on remote nodes to tackle multiple concurrent accesses or very-large data size) at run-time. D-NET offers a rich and expandable set of services targeting data collection, processing, storage, indexing, curation, and provision aspects. In the HOPE implementation, the D-NET toolkit is extended to include new services such as the Record Tagging and the Social Network Publishing and to adopt a "two-phase approach" to metadata records conversion. The ADI allows for the construction of an aggregated information space, populated by collecting (via OAI-PMH, HTTP, FTP) records from HOPE content providers and converting them

[1] http://www.loc.gov/marc/
[2] http://www.loc.gov/ead/
[3] International Association of Labour History Institutions, http://www.ialhi.org
[4] Europeana, http://www.europeana.eu

into a common HOPE format. Moreover, HOPE data curators can edit the aggregated records or tag them, in order to: (i) classify them, based on a vocabulary of historical themes defined by the consortium, or (ii) establish which social networks they should be sent to, based on a list of possible targets. Finally, the ADI makes the information space searchable and browsable by end-users from the project web portal and delivered to Europeana and other service consumers via OAI-PMH APIs.

The HOPE Common Data Model and XML Schema. The HOPE community comprises four "data provider profiles", namely library, archive, visual, audio video. Based on these, the project agreed on a common metadata model and its corresponding XML schema. In order to capture the commonalities of diverse object domains and formats, the model has been defined by studying the characteristics of the four profiles from the perspective of well-established standard formats in the respective field: MARCXML for libraries, EAD for archives, EN 15907[5] for audio video, and LIDO[6] for visual. The model includes seven classes of interrelated entities: Agent, Place, Event, Concept, Digital Resource, Theme, and Descriptive Unit (DU). DUs represent digital objects and include information about the real world object. According to the profiles, the DU class has four subclasses containing properties that are peculiar to one specific profile. Cross-domain properties are instead defined in the DU super class. DUs are related with each other via containment and sequential relationships so that it is possible to represents hierarchies of objects. A digital resource contains technical information about a digital representation of the object and is linked to the corresponding DU. Digital resources related to the same descriptive units can express sequential relationships, thus establishing a "reading path". Agents, places, concepts, events, and themes contextualize the object and are linked to DUs via relationships whose names describe the semantics of the association.

A "Two-Phase Approach" to Metadata Convertion. As pointed out by Haslhofer and Klas in [5], the use of mappings from each input format to the common format solves structural and semantic heterogeneities of metadata records, thus enabling the realization of homogeneous information spaces. In the case of HOPE, this process was complicated by the high degree of heterogeneity of input data sources: since the objects and metadata records collected from the content providers may belong to subcommunities of the overall ADI, the HOPE model tends to abstract over all of such communities and therefore the mapping from source models into the common model is not straightforward. For those reasons, the HOPE ADI implements a "two-phase approach". The first phase solves intra-profile structural and semantic heterogeneities, while the second phase solves inter-profile heterogeneities. The first phase is realized by mapping the metadata records of all data sources of the same profile onto metadata records conforming to a given standard data model for such profile; i.e. MARCXML (library), EAD (archive), EN 15907 (audio video), and LIDO (visual). The second phase is accomplished by providing mappings from such formats to the HOPE format.

[5] http://filmstandards.org/fsc/index.php/EN_15907

[6] http://network.icom.museum/cidoc/working-groups/data-harvesting-and-interchange/what-is-lido/

The approach brings two main benefits: it is easier for data source managers to map their formats into a standard format in their community; and the ADI can export data source content through standard formats without further data processing. On the other side, the adoption of standards can be a drawback for data richness in cases where the input format is richer than the adopted standard. For example, multilingual descriptions may be lost when mapping onto MARCXML. Once records are in the common format, their content is harmonized by applying vocabularies established by the consortium and compliant to standards (e.g., ISO country, ISO language). Moreover, curation and enrichment tools are available for data experts in order to: (i) check the quality of aggregated metadata record; (ii) create new virtual, cross-data source collections by tagging records with historical themes or social network publishing tags, e.g. objects tagged with “YouTube" are automatically exported to that social site. Finally, curated records are also transformed into the Europeana Data Model[7] (EDM) to be OAI-PMH harvested by Europeana. Social Network Publishing Services have also been deployed to react based on the aforementioned tagging actions.

Acknowledgements. This work is partly funded by the HOPE “Heritage of the People's Europe”, FP7 EU eContentplus, Best Practice Networks Project: Grant Agreement N. 250549. Its completion would have not been possible without the precious cooperation of the whole Project Consortium (http://www.peoplesheritage.eu/content/partners.htm).

References

1. HOPE Project, `http://www.peoplesheritage.eu`
2. eContentPlus framework, `http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm`
3. D-NET Software Toolkit, `http://www.d-net.research-infrastructures.eu`
4. Manghi, P., Mikulicic, M., Candela, L., Castelli, D., Pagano, P.: Realizing and Maintaining Aggregative Digital Library Systems: D-NET Software Toolkit and OAIster System. D-Lib Magazine 16(3/4) (2010)
5. Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Comput. Surv. 42(2), 7:1–7:37 (2010)
6. Bardi, A., Manghi, P., Zoppi, F.: Aggregative Data Infrastructures for the Cultural Heritage. In: Dodero, J.M., Palomo-Duarte, M., Karampiperis, P. (eds.) MTSR 2012. CCIS, vol. 343, pp. 239–251. Springer, Heidelberg (2012)

[7] `http://pro.europeana.eu/edm-documentation`

Towards a Methodology for Publishing Library Linked Data

Dydimus Zengenene[1], Vittore Casarosa[2], and Carlo Meghini[2]

[1] DILL International Master, University of Parma, Italy
[2] Istituto di Scienza e Tecnologie della Informazione del CNR, Pisa, Italy

Abstract. It is argued that linked data are becoming increasingly necessary for libraries and related institutions, such as galleries, museums and archives. Though libraries are potentially crucial players in the linked data movement, very often there is lack of knowledge among librarians on how to publish linked data. This paper presents the results of a master thesis whose main aim was to empower libraries to take an effective part in publishing linked data thereby contributing to building the semantic web, in order to improve the general services which they offer. In a narrower sense, this research aims to draw a methodology applicable to the library domain in publishing linked data. A 15-step methodology is presented and illustrated in some detail.

Keywords: linked data, library data.

1 Introduction

Due to their traditional role as the curators of valuable information, and their expertise in metadata generation and management, libraries or, more in general, the so called memory institutions (libraries, archives and museums) are in a unique position of providing trusted metadata for resources of cultural value. Libraries and related institutions are therefore expected or forced by circumstances to be key players in building the new generation of the web called the semantic web or the web of data. [2] notes that Libraries have already taken a leading role in the application of semantic web technologies because they own well described collections of objects.

Because of its ubiquity, scalability and simplicity, the web is recognized as the ideal medium to transform the way data is discovered, accessed, integrated and used. In that regard linked data can be defined as a set of principles and technologies that harness the ethos and infrastructure of the web to enable data sharing and reuse [8]. These concepts are penetrating also in the domain of Library and Information Science (LIS) and related disciplines, and [4] boldly states that "...of all information communities, libraries are in the best position to transition their data into Linked Data because the basic elements already exist in their catalogues".

The value of Linked Data for libraries derives mainly from the navigation possibilities offered by Linked Data, which provide a seamless information space that can be explored by following "typed" links (i.e. links with a known meaning) much in the

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 81–92, 2014.

same way as the “normal” web can be explored by following “anonymous” links. Links between libraries and non-library services such as Wikipedia, GeoNames, MusicBrainz, the BBC, and The New York Times will connect local collections into the larger universe of information on the Web. In addition, the use of globally unique identifiers (URIs) to designate works, places, people, events, subjects, and other objects or concepts of interest, allows libraries to increase their presence on the Web, by having their resources cited across a broad range of data sources. The use of unique identifiers also allows the description of a resource to be tailored to specific communities such as museums, archives, galleries, and audiovisual archives. Linked Data are represented in a structured way (RDF), which increases the visibility of library content towards crawling and relevancy algorithms of search engines and social networks.

Though libraries can potentially be crucial players in the linked data movement, very often there is lack of knowledge among librarians on how to publish linked data. The World Wide Web Consortium (W3C) has established a “Library Linked Data Incubator Group”, whose main aim was "to help increase global interoperability of library data on the Web, by bringing together people involved in Semantic Web...”. The final report of the group [12] points out that there are several reasons that make it a great challenge for libraries to successfully participate in the movement. One of the cited reasons is the complexity and variety of available vocabularies in libraries, their overlapping coverage, derivative relationships and alignments. The problem is exacerbated by the fact that most library and information professionals are unfamiliar with linked data sets and vocabularies that can be of use in the library domain because these data sets have been developed in the semantic web research community. There are efforts to participate in the movement, however most such library projects define themselves as prototypes.

The broad aim of this research is to empower libraries to take an effective part in publishing linked data thereby contributing to building the semantic web, in order to improve the general services which they offer. In a narrower sense, this research aims to draw a methodology applicable to the library domain in publishing linked data. In the remainder of this paper we provide e brief description of the basic principles underlying Linked Data (Section 2), then we describe briefly the methodology used to identify a 15-step process for publishing library data as linked data (Section 3), then we describe in some detail each of those steps (Section 4) and finally we provide some concluding remarks about the application of the process (Section 5).

2 Basics on RDF and Linked Data

Linked data is associated with a new generation of the web called the semantic web, the web of data or sometimes web 3.0. [8] define Linked Data as a set of “...best practices for publishing and connecting structured data on the Web”. The World Wide Web Consortium (W3C) shortly defines the semantic web as the web of data, which is capable of providing a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Linked Data is not about creating a different Web, but rather about enhancing the Web through the addition of structured information. This structured information, expressed using technologies such as URIs for identifiers of resources and RDF [14] for the description of their properties, specifies relationships between things, that can then be used for navigating between, or integrating, information from multiple sources.

RDF is based on the idea of identifying things (resources and their properties) using Web identifiers (called Uniform Resource Identifiers, or URIs), and describing resources in terms of simple statements asserting properties of a resource (sometimes called attributes) by providing a property and its and its associated values. This enables RDF to describe resources as a graph, where a node representing a resource (called the subject) can have a number of arcs, each one associated with a specific property, leading to another node (called the object) representing the value of that property. The value of a property (the object node) can be either a terminal string, in which case the graph ends there, or the URI of another resource, in which case the graph can continue with the properties if this new resource, which becomes the subject of a new set of statements.

In addition to providing a way to express simple statements about resources, using named properties and values, RDF also provides, through the RDF Schema, the ability for user communities to define the vocabularies (terms) that they intend to use in those statements: Those vocabularies (sometimes, maybe improperly, called ontologies) indicate the specific kinds or classes of resources that are going to be described, and the specific properties that are going to be used in describing those resources. RDF Schema provides the facilities needed to describe such classes and properties, and to indicate which classes and properties are expected to be used together (for example, to say that the property "jobTitle" will be used in describing a resource belonging to the class "Person").

In essence, Linked Data is a general way to organize information as "clouds of data" (RDF graphs) describing resources in specific application domains, which can be connected together in order to enrich each other and provide additional information. The resulting "global cloud" can be navigated to get the description of the resources and, following their properties and connections, to discover unexpected aspects or relationships of the objects of interest.

In the Linked Data world, it is important to distinguish between a resource, i.e. a real world object or concept, which not necessarily needs to be digital (and therefore downloadable from the Internet) and the description of that resource, which usually is digital and can be retrieved by navigating the Linked Data. It is common to provide the description of the resource in two ways, one expressed in HTML for use by a human through a browser, and one expressed in RDF/XML for use by an application e program. In this way, a resource usually leads to the definition of at least three URIs, the one that represents the real world object, the one that represents its HTML representation and the one that represents its RDF/XML representation.

3 Methodology

The data gathered by the W3C Incubator Group of Library Linked Data [12] was a starting point for this work. Even though this data was not collected to capture the

process, it was analysed identifying specific process issues that linked data publishers came across. These issues were used in conjunction with project reports to come up with questions which were used as a guideline for in-depth interviews. This approach was considered in order to avoid redoing the work that the W3C had already done but instead built on data which was collected by the W3C when it was preparing its report on Libraries and linked data. While W3C concentrated on the state of affairs in libraries, this work was concerned with implementation issues. This gave a different interpretation of the same data. However the data which was collected by the W3C was insufficient to meet the needs of this work since it was collected with a different intention. To cover for the shortfalls, we analyzed project reports of selected cases and prepared questions which were used for interviews. Books, opinion papers on linked data and web based resources were used to inform some sections of the work which both the data and the interviews could not fill. The complete results of the research are given in [13]. In this paper, we summarize the methodology, discussing its steps in some detail.

4 The Recipe

This section summarizes a 15 step practical methodology which can be followed for publishing linked data. As in libraries there are various datasets which can potentially be published as linked data, it should be clear that the exact workflow may vary depending on the nature of the data to be published, and also the fact that linked data is an evolving technology makes it difficult to define the "ideal workflow". However we can safely say that the library services and systems will become part of the "new web" if the library will manage to publish the following data: knowledge organization systems (classification schemes, thesauri), authority files, digital contents and their descriptions, catalogue data including circulation data sets. All these datasets should have links within themselves and should establish outgoing links to many other web resources, in order to attract many incoming links in what [3] calls "Web Centric Cataloguing".

4.1 Step 1: Motivation

It is important to know what is motivating a library to adopt linked data technologies. This implies understanding what linked data technology can do to improve your legacy system, how much it can improve users search and browsing experiences and how much it empowers them to contribute in enriching the information collection and in collaborating among themselves. The best motivations seem to come from comparing the current system against the potential of linked data enabled systems. Publishing the data on the web will make the organization more visible, and we may assume that this is what helps the library to better achieve its mission. This is supported by the fact that most libraries involved with large linked data projects are National Libraries and academic libraries, whose services may benefit by reaching a much wider public. All these considerations have to be written down clearly and convincingly in non technical jargon in order to use the document in the next step.

4.2 Step 2: Management Approval

Using the motivations in step one it is important to gain stakeholder approval and support. Management support to the project is important, given that so far no financial gains have been reported. On the contrary, some libraries have even abandoned their policy of selling metadata in favor of making them available for free as linked data. This is a management decision which one interviewee defined as "a great step". The benefits are also tangible since they can be measured by key library indicators, like user satisfaction, system improvement (precision, recall), innovative use and collaboration. This however depends on the ability for librarians to clearly view these benefits as a direct improvement of their services, clearly articulate the benefits and sell the idea to the stakeholders giving assurance on such sensitive issues like personal data security and controlled access to licensed resources.

4.3 Step 3: Sorting Out the Legal and Financial Issues

This stage has to do with assessing the rights that the institution has over the data sets. When there is a need to make contractual agreements with the data owners, there is also the need to know if the institution has the legal capacity to enter into such agreements. It is at this stage that license issues can be discussed, deciding what licenses and waivers the institution will grant to the intended users of the datasets. It is possible to have different licenses and waivers on different datasets of the same institution. The question of data licensing is still an open research issue [11]. Following linked data publishing principles does not necessarily mean opening up the data. There is an option for closed linked data. Miller and colleagues argue that the use of Creative Commons Licenses is mostly due to the lack of better alternatives, as pointed out at the site "Bibliographic Wilderness" (2008), which states that catalogue data is not "copyrightable".

4.4 Step 4: Assessment of Skills and Data Available

At this stage it is convenient to start planning the conversion process. In order to have a precise understanding of what is needed, this stage must be based on a situation analysis, which should include the assessment of the existing skills and the identification of the datasets to be published and the formats in which they are The options available have to be weighed against the skill that are needed. The fact that several cases reported to have relied on internal resources might suggest that it is not very difficult for motivated people to learn the needed skills "on the job". However, as noted by Coyle [4] librarians must transform their skills into understanding and managing ontologies, understanding information system design, etc., so that they can communicate with technical experts in carrying out such projects. Required skills:

- Information systems skills: these are of value in downloading, installing and configuring systems, databases (especially triple stores), and other servers, writing and reading XML and RDF.

- Metadata skills: knowledge of the cataloguing process and the meaning of the metadata elements.
- Modeling: understanding the data structure and coming up with the correct algorithms of converting data from a given structure to RDF.

If the library can hire external experts, the librarians need to communicate with the technical people in order to express their requirements. Therefore they need to:

- Be able to communicate their technical needs by speaking requirement modelling languages like the Universal Modeling Language.
- Be able to describe their ontologies using such vocabularies as FRBR, FRBRoo, CIDOC CRM, FOAF etc.
- Read and understand XML RDF records in various serialization formats to able to evaluate the end result.

Analyzing the data and the systems that hold them gives an overall idea of how complex the task might turn out to be. The diagram in Figure 1 gives an overall view of the conversion process, depending on the type of existing data. Depending on the type of data and the system where they are stored different conversion tools (such as entity extractors, RDF-izers, wrappers, RDB-to-RDF, etc, as indicated in the figure) should be used.

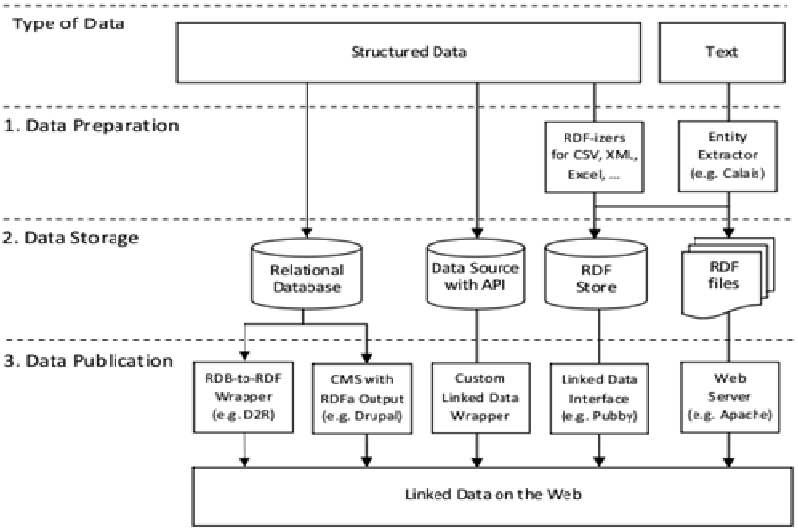


Fig. 1. Publishing workflow (from [8], p. 70)

4.5 Step 5: Tools Assessment and Evaluation

Referring again to Fig. 1, if the data exists in text format there is the need to use entity extractors, which provide RDF annotations on text documents. The resultant RDF

files can be published through a web server or through an RDF triple store. When the data is in other non queryable structured formats (e.g. spreadsheets), conversion takes place using other special tools referred to in the diagram as "RDF-izers" and then served via an RDF triple store. Many bibliographic systems store bibliographic data in queryable relational data format and digital libraries store their content in content management systems (CMS). In both cases there is little need to prepare the data, and for bibliographic data a custom wrapper should be used, while in the second case a CMS with RDFa output should be used. Knowing the data and the possible flow will help to know the skills, the experience and the tools that are needed. For example, using a custom wrapper is considered to be the most complex situation which cannot be accomplished without some programming skills [8].

4.6 Step 6: Dataset Analysis

This is a critical stage where most problems are likely to be faced. In addition to knowing the system that hosts the data and the data format, it is important to know also the level of description and the metadata elements sets in use. In the case of descriptions, very often they might be incomplete, or they might be based on systems that do not support linked data, or they might not follow any particular metadata standard. In the case of projects that involve aggregating data originating from different systems (e.g. from different institutions), or different systems like catalogues and digital library systems, there is the likelihood that these systems will be using different metadata standards (e.g. MARC and Dublin Core) and in this case there will be the need to map one set to the other.

4.7 Step 7: URI Assignment

Each resource in the dataset has to be identified by a unique URI, created according to the following guidelines:

- Use HTTP URIs so that they are dereferenceable.
- Ensure that the URIs are from a namespace that you control.
- Make sure your URIs do not carry implementation details which can change over time.
- It is advisable to use meaningful natural keys in URIs as unique identifiers of resources; for example, books can be identified by using the ISBN number instead of primary keys in the local database.

As described before, one resource in a dataset usually leads to the creation of at least three URIs, the one that represents the real world object, the one that represents its HTML representation and the one that represents its RDF/XML representation. In the world of Linked Data there are three common ways to distinguish these URIs. The most common one, used by dbpedia, Europeana and others, is the use of the terms "resource", "page" and "data" in their URIs, to indicate the real world object, its HTML representation and its RDF/XML representation respectively [8]. Another alternative is to use the terms "id","pages" and "data" on their URIs. Finally, the third

alternative is to use the file extensions, using no extension for the resource, "html" for the HTML representation and "rdf" for the RDF representation.

Whereas making URIs for resources that you control involves making a choice of the representation that you prefer, external URIs do not give such freedom. In many cases, different institutions refer to the same resource using different URIs. If you have the same resource in your data set, which URI should you use? One alternative is to define your own, and then declare its equivalency with the others by using the "owl:sameAs" property. Another alternative is to use an existing one, taking into consideration the authority and reliability of the institution that defined it. For domain specific datasets one may find domain specific authorities like GeoNames for geographic names, VIAF for author names in libraries, FOAF for people etc. For general datasets more general authorities like dbpedia may be more useful.

4.8 Step 8: Vocabulary Modeling

Vocabulary modelling has to do with creating controlled terminology giving explicit meaning to the concepts in your dataset, and this is a key process in linked data. The emphasis is on the use of already existing vocabularies evaluated using the four criteria as noted by [8] (p.62). A vocabulary of choice should:

- be widely used to ensure widespread use of your dataset,
- be actively maintained according to a clear governance process,
- cover enough of your dataset to justify its terms, and
- be expressive enough to suit your particular requirements.

There is no single place to find vocabularies. From the literature, libraries are recommended to consider their own domain models as a starting point. These include FRBR, FRAD, and the more recent FRBRoo [4]. Alternatively, it is recommended to find suitable vocabularies by asking what other people in the same domain are using. Swoogle, Sindices and Taxonomy warehouse are some of the sources to find vocabularies [1]. Also the W3C supplementary report on data sets, vocabularies and metadata elements sets is a good starting point to see what others have used before.

A good knowledge of your existing data will definitely help in the choice of a vocabulary. Does your data include People, Places, Books, Journals, Films, Authors, Musicians, Concepts, Photos, Comments, Reviews and so on. This knowledge makes it possible to have quick choices of vocabularies which are widely used at the moment. For example, Geonames would be among the first to consider for Places, and FOAF would be the one to consider for People [7]. In the most common cases some new terms are introduced as different existing vocabularies are brought together to satisfy one's need. Such new terms are called proprietary terms. They should be made deferenceable to ensure that applications and other users can retrieve their definitions in order to know their meaning and perhaps reuse them. In the worst case one might be forced to build a new vocabulary from scratch. One major reason why it is always preferable to reuse existing vocabularies is the fact that whenever you create a new vocabulary in the linked data world you have created a data island and have decreased the level of understanding in that domain. Unless you are a very well known authority in that domain, it is likely that your vocabulary will not be used by someone else.

Creating a vocabulary is a very complex process which needs linguistic skills and domain expertise. Creating your own vocabulary should be the very last option in vocabulary modelling. In the event that you create your own vocabularies, consider relating your new vocabulary to known vocabularies making them sub concepts of the known vocabularies.

4.9 Step 9: Generation of RDF Data

If the preceding steps were done well, the process of generating the RDF data should not be a difficult one, as this is the stage of implementation of decisions taken before. At the moment there is a number of tools being developed to help in this process, so that in the future there might be less need to have strong coding skills but rather skills for configuring and customizing would be needed. In many cases, the amount of data is too large to convert existing literal values into URIs manually, and the use of tools to do the process automatically or semi-automatically is mandatory. It is not part of this work to compare and contrast existing tools (some of them are being developed in house by major projects) as each of them has its own strengths and weaknesses. Reviews of these systems can be found easily on the web.

4.10 Step 10: Enriching the Data

This process involves defining news triples that connect resources within the published dataset or to other resources in other external datasets, creating triples that define relationships internally within the dataset (internal links) or relationships with outside resources (outgoing links). For internal links it must be ensured that every part of the dataset is reachable by a crawler when it is following links and therefore each file has to be connected to related files in the same dataset. For outgoing links, it is advisable to start by linking to such datasets as dbpedia, Geonames, Europeana, VIAF and others, which are already well established and stable in the linked data world. This ensures that your dataset is easily discoverable since these are widely linked to by many other datasets. As already stated, points of caution are: (1) URIs must not carry implementation detail which can change over time (e.g., port numbers, server names or php extensions). (2) Use natural keys in URIs as unique identifies of resources (e.g., ISBN number instead of using the primary key in your local database). (3) natural key should be meaningful within the domain (e.g. ISBN is meaningful in the library domain). For external links ensure that you are connecting to datasets that adds value to your set and are reliable. The questions to be asked when linking to external datasets are:

- What is the value of your data in the target dataset?
- To what extent does this add value to the new dataset?
- Is the target dataset and its namespace under stable ownership and maintenance?
- Are the URIs in the dataset stable and not likely to change?
- Are there outgoing links to other datasets?

There is research going on about ranking datasets. Such works include “DING”, a novel two-layer ranking model for the Web of Data [5].

4.11 Step 11: Describing the Data Set

Before publishing, there is the need to provide a description of the dataset, which includes provenance and licensing metadata. In provenance metadata one might describe the history of that dataset, how it was generated and the technical processes that have been undergone to establish the dataset. In license and waiver metadata, one describes how that dataset maybe used by third parties. A good description of the dataset is therefore essential to establish trust among third parties and also to ensure that people will use the dataset according to the conditions set by the publisher.

In addition to making instance data self-descriptive, it is also desirable that data publishers provide metadata describing characteristic of complete data sets, for instance, the topic of a data set and, if possible, more detailed statistics about the data. This information might include label, URI of the dataset, location of SPARQL endpoint, data dumps, last-modified date of the dataset, and change frequency [8] (p. 48). These data may be described using the Vocabulary of Interlinked datasets (voiD), which is an RDF vocabulary, as in the "datahub" web site [9].

4.12 Step 12: Evaluating the Dataset

Before publishing the data for access by third parties, it is advisable to evaluate it to see how good it is and if it conforms to standards. A possible checklist proposed by [8] (p. 83) is the following.

- Does your data set links to other data sets?
- Do you provide provenance metadata?
- Do you provide licensing metadata?
- Do you use terms from widely deployed vocabularies?
- Are the URIs of proprietary vocabulary terms dereferenceable?
- Do you map proprietary vocabulary terms to other vocabularies?
- Do you provide data set-level metadata?
- Do you refer to additional access methods?

At this stage is also advisable to check if the triples are well expressed. This can be done by using such tools as Vapour Linked Data Validator, RDF: Alerts and Sindice Inspector. In addition, there is another level of evaluating a dataset, which involves evaluating the data quality. This has to do with how you can find the goodness of a dataset and weather it is worth linking to. Some key criteria have been mentioned in the step of selecting a vocabulary to link to. Finally, it is also advisable to use the 5 stars rating system recommended by W3C.

4.13 Step 13: Publishing

The decision of how the data will be published should have been done in Step two, where the workflow was planned, and where the need to provide users with several access points to ensure that the dataset is widely used should have been considered. The most obvious way to publish Linked Data on the Web is to make the URIs that identify data items dereferenceable into RDF descriptions. In addition, various Linked

Open Data providers, including libraries, provide two alternative means of accessing the data, namely via SPARQL endpoints or by providing RDF dumps of the complete data set. In general the system should provide access to both the RDF and HTML representations of the data. This is usually done by configuring 303 redirects in triple stores in response to a client request to access either the HTML representation of an object or its RDF representation.

4.14 Step 14: Incoming Links

Incoming links originate from other datasets, linking into your dataset. Third parties need to be convinced that your dataset is valuable to them so that they can link to it. However it is usually difficult for them to know your value unless you do some sort of marketing and promotion actions. As a starting point a publisher can create triples that link to their own dataset and ask third parties like dbpedia to add those triples to their own dataset [8] (p.64) To continue attracting new links, there might be the need to employ marketing techniques so that the dataset is known by new users and be linked to.

4.15 Step 15: Feedback

Once the data has been published and incoming links solicited, it is important to wait for feedback from the user community so that one can incorporate their views and refine the dataset. From the interviews conducted it was found out that this is a continuous process, even if it is usually very difficult, if not impossible, to know what people are doing with the RDF data. Download counts could give a rough idea of your data usage, and Google Analytics could provide additional information, such as the change in the number of times a certain URI is accessed, the IP addresses which are accessing your data and their geographic location, so that one could derive a rough idea of how the data is being used.

5 Conclusions

Changes brought about by the internet and the web have affected all walks of life and changed business models under which many institutions have been operating. Libraries and the library profession have not been spared. In order to continue being relevant and effective, libraries have to adapt to the latest developments in technology and the resultant user behaviour. Adopting linked data technologies is one such initiative to improve the library presence where today's information is sought (i.e. the web), to improve the functionality of their systems and to promote innovative use of their data. Even though the 15 step approach helps in following the progress and managing the project, it is recommended taking first a small dataset and carrying out the necessary trials and errors until the data can be successfully published, in order to develop the needed skills. Clearly, for large datasets and large projects that need to be monitored and evaluated, planning and accountability tracing is necessary thereby calling for a methodological approach. This procedural approach should hopefully minimize possible mistakes.

References

1. Where to find vocabularies, http://vocamp.org/wiki/Where_to_find_vocabularies
2. Isaac, A., Kramer, D., van der Meij, L., Wang, S., Schlobach, S., Stapel, J.: Vocabulary matching for book indexing suggestion in linked libraries – A prototype implementation and evaluation. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 843–859. Springer, Heidelberg (2009)
3. Chudnov, D.: What linked data is missing. Computers in Libraries 31(8) (2011)
4. Coyle, K.: Making connections. Library Journal 134(7), 44–47 (2009)
5. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G., Decker, S.: Hierarchical link analysis for ranking web data. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010, Part II. LNCS, vol. 6089, pp. 225–239. Springer, Heidelberg (2010)
6. Europeana Foundation. The problem of the yellow milkmaid: A business model perspective on open metadata (2011), http://pro.europeana.eu/docu-ments/858566/2cbf1f78-e036-4088-af25-94684ff90dc5/
7. Heath, T.: An introduction to linked data (2009), http://tomheath.com/slides/2009-02-austin-linkeddata-tutorial.pdf
8. Heath, T., Bizer, C.: Linked data. Evolving the web into a global data space. Synthesis Lectures on the Semantic Web. Morgan & Claypool (2011)
9. Taxonconcept knowledge base, http://datahub.io/
10. Jackson, J.: Open source vs. proprietary software. PCWorld (2011)
11. Heath, R., Miller, T., Styles, P.: Open data commons, a license for open data. In: Proc. of LDOW 2008-Linked Data on the Web Workshops, Beijing, China (2008)
12. Coyle, K., Dunsire, G., Isaac, A., Murray, P., Panzer, M., Schneider, J., Singer, R., Summers, E., Waites, W., Young, J., Zeng, M., Baker, T., Bermes, E.: Library linked data incubator group final report. W3C (October 2011), http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/
13. Zengenene, D.: A methodology for publishing linked data in the library domain. Master's thesis, International Master in Digital Library Learning (2012)
14. Manola, F., Miller, E.: RDF Primer. Technical report, W3C (February 2004), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

ConNeKTion: A Tool for Handling Conceptual Graphs Automatically Extracted from Text

Fabio Leuzzi[1], Stefano Ferilli[1,2], and Fulvio Rotella[1]

[1] Dipartimento di Informatica, Università di Bari
{fabio.leuzzi,stefano.ferilli,fulvio.rotella}@uniba.it
[2] Centro Interdipartimentale per la Logica e sue Applicazioni, Università di Bari

Abstract. Studying, understanding and exploiting the content of a digital library, and extracting useful information thereof, require automatic techniques that can effectively support the users. To this aim, a relevant role can be played by concept taxonomies. Unfortunately, the availability of such a kind of resources is limited, and their manual building and maintenance are costly and error-prone. This work presents ConNeKTion, a tool for conceptual graph learning and exploitation. It allows to learn conceptual graphs from plain text and to enrich them by finding concept generalizations. The resulting graph can be used for several purposes: finding relationships between concepts (if any), filtering the concepts from a particular perspective, extracting keyword, retrieving information and identifying the author. ConNeKTion provides also a suitable control panel, to comfortably carry out these activities.

1 Introduction

The spread of the electronic technology has had a dramatic impact on the production of documents in all fields of knowledge, and has led to the flourishing of document repositories aimed at supporting scholars and non-technically aware users in carrying out their tasks and satisfying their information needs. However the study, understanding and exploitation of the content of a digital library, and the extraction of useful information thereof, are complex activities requiring automatic techniques that can effectively support the users. In this landscape, a relevant role can be played by concept taxonomies that express both common sense and domain-specific information, including implicit relationships among the concepts underlying the collection. Unfortunately, the availability of such a kind of resources is limited, and their manual building and maintenance are costly and error-prone. A possible solution is the exploitation of Natural Language Processing (NLP) techniques, that work on the textual parts of the documents to extract the concepts and relationships expressed by words. Although the task is not trivial, due to the intrinsic ambiguity of natural language and to the huge amount of required common sense and linguistic/conceptual background knowledge, even small portions of such a knowledge may significantly improve understanding performance, at least in limited domains. This work presents ConNeKTion (acronym for 'CONcept NEtwork for Knowledge representaTION'), a

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 93–104, 2014.

tool for conceptual graph learning and exploitation. It allows to learn conceptual graphs[1] from plain text and to enrich them by finding concept generalizations. The resulting graph can be used for several purposes: finding relationships between concepts (if any), filtering the concepts from a particular perspective, keyword extraction, information retrieval and author identification. Specifically, this paper focuses on the graphical control panel provided to the user for exploiting the various functionalities, while technical details and evaluation of the single functionalities have been already presented in [10, 16, 24].

The paper is organized as follows: the next section describes related works that have some connection to the present proposal, described in Section 3; then, Section 4 describes the tool aimed at supporting end users in managing and exploiting the learned conceptual graph. In the last section, some considerations and future work issues are proposed.

2 Related Work

A primary task in this work is the construction of a conceptual graph starting from the analysis of the plain text contained in the documents that make up a collection. Several techniques are present in the literature generally aimed at building some kind of graph-like structure, that have made the basis on which the state of the art specifically aimed at building taxonomies and ontologies from text has built its approaches. [2] builds concept hierarchies using Formal Concept Analysis: objects are grouped using algebraic techniques based on their descriptive attributes, which are determined from text linking terms with verbs. Different approaches are also available. [18, 17] build ontologies by labeling taxonomic relations only; in our opinion, also non-taxonomic relationships are very important to improve text understanding, such as those associated to actions (and expressed by verbs). [21] builds taxonomies considering only concepts that are present in a domain but do not appear in others; however, one might be interested in collecting and organizing all concepts that can be recognized in a collection, because generic ones may help to frame and connect domain-specific ones. [20] defines a language to build formal ontologies, but this level is very hard to be effectively reached, so a sensible trade-off between expressive power and practical feasibility might better focus on working at the lexical level (at least in the current state of the art).

Our proposal to learning conceptual graphs from text relies on pre-processing techniques taken from the field of NLP, that provide a formal and structured representation of the sentences on which the actual graph construction and reasoning operators can be applied. As regards the syntactic analysis of the input text, the *Stanford Parser* and *Stanford Dependencies* [15, 3] are two well-known tools that can be trained for any language for the identification of the most likely syntactic structure of sentences (including active/passive and positive/negative forms), and specifically their 'subject' or '(direct/indirect) object' components.

[1] We use the term 'conceptual graph' as a synonym for 'concept network', with no reference to Sowa's formalism.

They also normalize the words in the input text using lemmatization instead of stemming in order to preserve the grammatical role of the original word (and improve readability by humans). Due to the wide range of tools available for English, compared to other languages, we will focus on this language in the following.

Also, we need in some steps of our technique to assess the similarity among concepts in a given conceptual taxonomy. A classical, general measure, is the *Hamming distance* [11], that works on pairs of equal-length vectorial descriptions and counts the number of changes required to turn one into the other. Other measures, specific for conceptual taxonomies, are [9] (that adopts a global approach based on the whole set of super-concepts) and [31] (that focuses on a particular path between the nodes to be compared).

Another technology we use is ProbLog [22] to apply probabilistic reasoning on the extracted knowledge. It is essentially Prolog where all clauses are labeled with the probability that they are true, that in turn can be extracted from large databases by various techniques. A ProbLog program $T = \{p1 : c1, ..., pn : cn\}$ specifies a probability distribution over all its possible non-probabilistic subprograms according to the theoretical basis in [28]. The semantics of ProbLog is then defined by the success probability of a query, which corresponds to the probability that it succeeds in a randomly sampled program. Indeed, the program can be split into a set of labeled facts $p_i :: f_i$, meaning that f_i is a fact with probability of occurrence p_i, and a Prolog program using those facts, which encodes the background knowledge (BK). Probabilistic facts correspond to mutually independent random variables, which together define a probability distribution over all ground logic programs $L \subseteq L_T$ (where L_T is the set of all f_i's):

$$P(L|T) = \prod_{f_i \in L} p_i \prod_{f_i \in L_T \setminus L} (1 - p_i)$$

In this setting we will use the term *possible world* to denote the least Herbrand model of a subprogram L together with BK, and we will denote by L both the set of sampled facts and the corresponding world.

A possible exploitation of the learned conceptual graph is for Information Retrieval (IR) purposes, so a quick overview of this field of research may be useful as well. Most existing works that tackle the IR problem are based on the so-called *Vector Space Model* (VSM), originally proposed in [27]. This approach represents a corpus of documents D, and the set of terms T appearing therein, as a $T \times D$ matrix, in which the (i, j)-th cell reports a weight representing the importance of the i-th term in the j-th document (usually computed according to both the number of its occurrences in that document and its distribution in the whole collection). Many similarity approaches [13, 26] and weighting schemes [25, 23, 29] have been proposed. Based on this representation, the degree of similarity of a user query to any document in the collection can be computed, simply using any geometric distance measure (e.g., the cosine measure) on that space. One limitation of these approaches is their considering a document only from a lexical point of view, which is typically affected by several kinds of linguistic tricks, such

as synonymy and polysemy. More recently, techniques based on dimensionality reduction have been explored with the aim to map both the documents in the corpus and the queries into a lower dimensional space that explicitly takes into account the dependencies between terms, in order to improve the retrieval or categorization performance. Prominent examples are Latent Semantic Indexing [4] (a statistical method based on Singular Value Decomposition that is capable of retrieving texts based on the concepts they contain, not just by matching terms) and Concept Indexing [14] (that exploits Concept Decomposition [5] instead of Singular Value Decomposition).

3 Provided Functionalities

ConNeKTion aims at partially simulating some human abilities in the text understanding and concept formation activity, such as: extracting the concepts expressed in given texts and assessing their relevance; obtaining a practical description of the concepts underlying the terms, which in turn would allow to generalize concepts having similar descriptions; applying some kind of reasoning 'by association', that looks for possible indirect connections between two identified concepts; identifying relevant keywords that are present in the text and helping the user in the retrieval of useful information; building a model of the author's writing style through a relational description of the syntactical structure of the sentences, in order to understand whether two documents have been written by the same author or not. The system takes as input texts in natural language, and process them to build a conceptual network that supports the above objectives. The resulting network can be considered as an intensional representation of a collection of documents. Translating it into a suitable First-Order Logic (FOL) formalism allows the subsequent exploitation of logic inference engines in applications that use that knowledge.

3.1 Graph Learning

ConNeKTion exploits a mix of existing tools and techniques, that are brought to cooperation in order to reach the above objectives, extended and supported by novel techniques when needed.

Natural language texts are processed by the Stanford Parser in order to extract triples of the form $\langle subject, verb, complement \rangle$, that will represent the concepts (the *subject*s and *complement*s) and relationships (*verb*s) for the graph. Some representational tricks are adopted: indirect complements are treated as direct ones by embedding the corresponding preposition into the verb; sentences involving verb 'to be' or nouns with adjectives contributed in building the sub-class structure of the taxonomy (e.g., "the penguin is a bird" yields $is\_a(penguin,bird)$). Specifically, 'is_a' relationships are exploited to build the taxonomy. The representation formalism was enriched by including the sentence's positive or negative form based on the absence or presence (respectively) of a *negation modifier* for the verb in the corresponding syntactic tree. The frequency of each arc between the concepts in positive and negative sentences were

taken into account separately. This made our solution more robust, laying the basis for a statistical approach that inspects the obtained taxonomy by filtering out all portions that do not pass a given level of reliability.

The graph so built embeds formal descriptions of concepts, on which the use of generalizations provides many opportunities of enrichment and/or manipulations on the graph. It can be used to build taxonomic structures, also after the addition of new text (possibly causing the presence of new nodes in the graph); to shift the representation, by removing the generalized nodes from the graph and leaving just their generalization (that inherits all their relationships to other concepts); to extend the amount of relationships between concepts belonging to the same connected component of the graph, or to build bridges between disjoint components that open new reasoning paths (which improves the effectiveness of reasoning 'by association'). Given two concepts G and C, G generalizes C if anything that can be labeled as C can be labeled as G as well, but not *vice-versa* [16]. The generalization procedure is made up of three steps: *Concept Grouping*, in which all concepts are grossly partitioned to obtain subsets of concepts (we group similar concepts if the aim is to enrich the relationships, or dissimilar ones in the bridging perspective); *Word Sense Disambiguation*, that associates a single meaning to each term by solving possible ambiguities using the domain of discourse; *Computation of taxonomic similarity*, in which WordNet [7] is exploited in order to further filter with an external source the groups found in step 1, and to choose an appropriate subsumer.

3.2 Reasoning by Association

We intend 'reasoning by association' in a given conceptual graph as the task of finding a path of pairwise related concepts that establishes an indirect interaction between two concepts [16]. Our tool provides two different strategies for doing this: one works in breadth and returns the minimal path (in the number of traversed edges) between concepts, also specifying all involved relations; the other works in depth and allows to answer probabilistic queries on the conceptual graph.

In more details, the former strategy looks for a minimal path starting two *Breadth-First Search* (BFS) procedures, one for each concept under consideration, until their boundaries meet. It also provides the number of positive/negative instances, and the corresponding ratios over the total, in order to express different gradations (such as permitted, prohibited, typical, rare, etc.) of actions between two objects. While this value does not affect the reasoning strategy, it allows to distinguish which reasoning path is more suitable for a given task. Note that this is different than the standard *spreading activation* algorithm, in that (1) we do not impose weights on arcs (we just associate arcs with symbolic labels expressing their semantics) nor any threshold for graph traversal, (2) we focus on paths rather than nodes, and specifically we are interested in the path(s) between two particular nodes rather than in the whole graph activation, hence (3) it makes no sense in our approach setting the initial activation weight of start nodes, and (4) this allows us to exploit a bi-directional partial search rather than a mono-directional complete graph traversal.

Since real world data are typically noisy and uncertain, the latter strategy was included, that softens the classical rigid logical reasoning. This is obtained by suitably weighting the arcs/relationships among concepts to represent their likelihood among all *possible worlds*, and using these weights to prefer some paths over others. ProbLog [22] is exploited for this purpose, whose descriptions are based on the formalism $p_i :: f_i$ where f_i is a ground literal having probability p_i. In our case, f_i is of the form *link*(*subject*, *verb*, *complement*) and p_i is the ratio between the sum of all examples for which f_i holds and the sum of all possible links between *subject* and *complement*. Again, this is different than *spreading activation* because the ProbLog strategy is adopted.

3.3 Keyword Extraction

The identification of relevant nodes in the graph may in some sense correspond to selecting keywords that provide indications on the main topics treated in the collection. As a first step, the frequency of each term is computed (its spread through the collection is ignored, to allow the incremental addition of new texts without the need of recomputing this statistics). Then, the EM clustering approach provided by Weka based on the Euclidean distance is applied to row vectors (representing concepts in the graph). Finally, various Keyword Extraction techniques, based on different (and complementary) aspects, perspectives and theoretical principles, are applied on the input texts to identify relevant concepts. We mixed a quantitative approach based on co-occurrences [19], a qualitative one based on WordNet [8] and a novel psychological one based on word positions. Assuming that humans tend to place relevant terms/concepts toward the start and end of sentences and discourses, where the attention of the reader/listener is higher [12], this approach determines the chance of a term being a keyword based on its position in the sentence/discourse. In particular, a mixture model determined by two Gaussian curves, whose peaks are placed around the extremes of the portion of text to be examined, is used. The outcomes of these techniques are exploited to compute a compound *Relevance Weight* for each node in the network. Then, nodes are ranked by decreasing Relevance Weight, and a suitable cut-point in the ranking is determined to distinguish relevant concepts from irrelevant ones. We cut the list at the first item in the ranking such that the difference in relevance weight from the next item is greater or equal than the maximum difference between all pairs of adjacent items, smoothed by a user-defined parameter $p \in [0, 1]$ [10].

3.4 Information Retrieval

The set of obtained representative keywords for each document can be considered as a higher-level representation of the digital library's content, and hence keyword extraction also work as a pre-processing step toward Information Retrieval in the library itself [24]. Indeed, to each keyword a corresponding meaning can be associated as follows: each keyword in the document is mapped to a corresponding synset (i.e., the code of a concept) in WordNet, that is taken as its semantic

representative, using Word Sense Disambiguation techniques [10]. The output such a step, for each document, is a list of pairs, consisting of keywords and their associated synsets. All these synsets are partitioned into different groups using pairwise clustering. Then, each document is considered in turn, and each of its keywords 'votes' for the cluster to which the associated synset has been assigned. The aim is finding groups of similar synsets that might be usefully exploited as a kind of 'glue' binding together subsets of documents that are consistent with each other. In this perspective, the obtained clusters can be interpreted as intensional representations of specific domains, and thus they can be exploited to retrieve the sub-collection they are associated to. In this setting, a query in natural language is processed in order to recognize the relevant terms, and consequently find the corresponding synsets. At this point, a similarity evaluation (using the function in [8]) is performed against each cluster (that has a list of associated documents). The best result is used to obtain the list of documents by descending relevance, that can be used as an answer to the user's search.

3.5 Author Identification

This functionality wants to face a well-known problem [1, 6, 30]: given a set of documents by a single author and a questioned document, determine whether the questioned document was written by that particular author or not.

This technique is based on First-Order Logic. It is motivated by the assumption that making explicit the typed syntactical dependencies in the text one may obtain significant features on which basing the predictions. Thus, this approach translates the complex data represented by natural language text to complex (relational) patterns that allow to model the writing style of an author.

Our approach consists in translating the sentences into relational descriptions, then clustering these descriptions (using an automatically computed threshold to stop the clustering procedure). The resulting clusters represent our model of an author. So, after building the models of the base (known) author and the target (unknown) one, the comparison of these models suggests a classification (i.e., whether the target author is the same as the base one or not). The underlying idea is that the model describes a set of ways in which an author composes the sentences in its writings. If we can bring back such writing habits from the target model to the base model, we can conclude that the author is the same.

4 Exploitation Tool

The above functionalities are delivered to the users through a graphical tool that provides a set of controls allowing to explore and analyze the conceptual graph. The learned net is represented through an XML file. The tool can load a file in this format, and draw the corresponding net (automatically organizing the nodes in the best possible way). The semantic network can be built incrementally, avoiding too long times of unavailability. Different colors are used for nodes

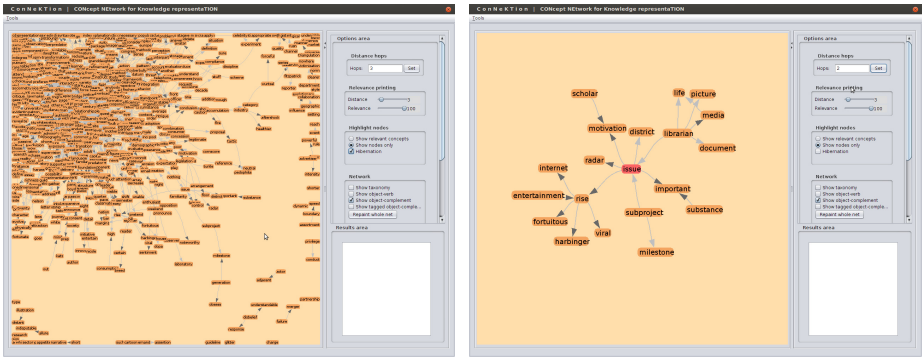


Fig. 1. The main control panel

depending on their type: subjects and complements have a different color than verbs. Also the relations are filled with a different color depending on the positive or negative valence of the corresponding phrase. Figure 1 shows two screenshots of the main interface of the tool, showing two different perspectives on the same net (a complete overview and a selection thereof, respectively). The main area, on the left, contains the net.

After loading a conceptual graph, the tool allows to explore it in a graphical intuitive way, using classical mouse-based controls. Since the compound view of the whole graph is typically cluttered and very dense of (often overlapping) nodes and edges (but still useful to grasp the overall shape of the net), scroll, pan and zoom in/out controls allow to focus on specific parts thereof and to have a better insight on them. Single nodes can be dragged as well, and the entire net is automatically rearranged accordingly to best fit the available space. Finally, by selecting a specific node, it is possible to set a neighborhood limit such that all the nodes whose shortest path to the selected node are outside the selected level are filtered out.

All the controls, settings and results are placed in a panel standing on the right of the graph visualization window. Such a panel is in turn divided into several sub-parts (shown in Figures 2, 3 and 4). Let us examine the single sub-areas of the control panel in more details.

Distance hops is in the top part of the panel (shown in Figure 3) and containing a text field in which the user can enter the desired level up to which nodes are to be shown, starting from the selected one (i.e. the user can set the maximum neighborhood level). **Relevance filtering** contains two sliding bars: the former is *Distance*, it is aimed at providing the same functionality as the *Distance hops* area, but bound in $[0, maxHops(net)]$, where $maxHops(\cdot)$ is a function that returns the diameter of a given net; the latter is *Relevance*, it allows to set the relevance parameter that determines which relevant nodes are to be highlighted. **Highlight nodes** is a radio button, it allows to select a visualization that highlights relevant nodes or a classical one. Choosing the relevant nodes perspective enables access to the advanced functionality of relevant node

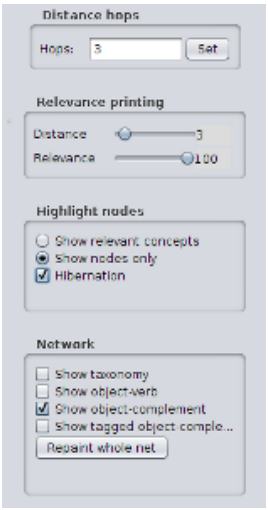


Fig. 3. Set of parameters placed in the right panel

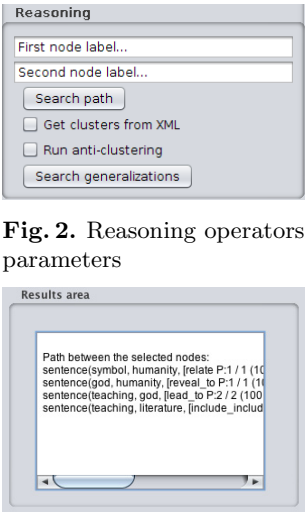


Fig. 2. Reasoning operators parameters

Fig. 4. Textual area dedicated to results

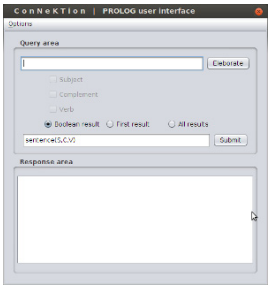


Fig. 5. The Prolog KB query tool

recognition, useful for a deeper analysis of the collection. Moreover, using the *Hibernation* check box the study of the net is made more comfortable for the human reader. This issue may require further clarification. In standard mode, the research for a balanced placement of nodes within the used space is always 'alive', so that the nodes automatically rearrange their position in the screen after the perturbations introduced by the user when he moves some elements to study them more comfortably. Since the continuous movement of the nodes makes the visual analysis of the net difficult, the possibility to stop the net in order to explore it (through reading and manual rearrangements of single nodes) was introduced. **Network** embeds four options, and specifically: *Show taxonomy*, that adds taxonomic relations to the network; *Show object-verb*, that adds verbs as nodes, and edges $< subject, verb >$ and $< verb, complement >$; *Show object-complement*, that adds direct relations $< subject, complement >$ (regardless of the verbs connecting them); *Show tagged object-complement*, that enables the tagging of the relations $< subject, complement >$ with verbs and associated (positive or negative) valence as reported in the XML file (so, the visual outcome is the same as for *Show object-complement*, but the tagged relations in the XML can be used for further functionalities). **Reasoning** is devoted to the reasoning operators (shown in Figure 2). In particular, it contains two text fields in each of which a concept (label of a node) can be entered, so that pressing the *Search path* button starts the Reasoning by Association functionality in order to obtain a plausible complex relation between the specified concepts. This sub-area also contains a button (named *Search generalizations*) that starts the search for *Generalization*; its behavior can be modified by acting on two checkboxes, *Get clusters from XML* (that avoids computing the clusters if they have already been

computed and stored in suitable XML files), and *Run anti-clustering* (that starts the technique to build bridges between different components of the net [16]). **Results** appears in the bottom area in the panel (shown in Figure 4). It is dedicated to textual results, consisting of paths where each row reports in square brackets (the labels of) the relations that exist between two nodes. In particular, each relation (verb) is associated to the number of positive and negative instances in which it occurred, expressing its valence. This also provides an indication of the degree of reliability of the path sub-parts. As an example, the screenshot in Figure 4 shows the resulting path between nodes 'symbol' and 'literature':

sentence(symbol, humanity, [relate P: 1/1 (100.0%), N: 0/1 (0.0%)])
sentence(god, humanity, [reveal_to P: 1/1 (100.0%), N: 0/1 (0.0%)])
sentence(teaching, god, [lead_to P: 2/2 (100.0%), N: 0/2 (0.0%)])
sentence(teaching, literature, [include_including P: 4/4 (100.0%), N: 0/4 (0.0%)])

which can be interpreted as: "Humanity can relate by means of symbols. God reveals to humanity. Teaching (or education) leads to God, and includes the use of literature.". Here only one relation per row is present, and there are no sentences with negative valence.

Finally, an additional functionality concerns the possibility of querying the ProLog knowledge base expressing the content of the net, which allows more complex kinds of reasoning than simple reasoning by association on the graph. It can be accessed from menu *Tools* in the main window, using the *PROLOG user interface* item. A window like that in Figure 5 is opened, that allows to enter a ProLog query to be answered using the knowledge base (e.g., "what does a dog eat?" might be asked in the form *eat(dog,X)*). The ProLog representation of the net can be obtained and saved from the same window, by choosing the *Create new K.B.* item in the *Options* menu.

5 Conclusions

Studying, understanding and exploiting the content of a digital library, and extracting useful information thereof, are complex and knowledge-intensive activities for which the user needs the support of effective automatic techniques. To this aim, a relevant role can be played by concept taxonomies. Unfortunately, the availability of such a kind of resources is limited, and their manual building and maintenance are costly and error-prone. ConNeKTion is a tool that allows to learn conceptual graphs from plain text and to enrich them by finding concept generalizations. The resulting graph can be used for several purposes: finding relationships between concepts (if any), filtering the concepts from a particular perspective, keyword extraction, information retrieval and author identification. A suitable control panel is provided for the user to comfortably carry out these activities.

As future work, we plan to improve the natural language text pre-processing using anaphora resolution in order to replace, where possible, pronouns with the explicit concept they express. We also wish to extend the reasoning operators by adding an argumentation operator, that could exploit probabilistic weights, intended as a rate of reliability, to provide support or attack to a given statement.

References

[1] Argamon, S., Saric, M., Stein, S.S.: Style mining of electronic messages for multiple authorship discrimination: first results. In: Getoor, L., Senator, T.E., Domingos, P., Faloutsos, C. (eds.) KDD 2003, pp. 475–480. ACM (2003)

[2] Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. J. Artif. Int. Res. 24(1), 305–339 (2005)

[3] de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure trees. In: LREC (2006)

[4] Deerwester, S.: Improving Information Retrieval with Latent Semantic Indexing. In: Borgman, C.L., Pai, E.Y.H. (eds.) Proceedings of the 51st ASIS Annual Meeting (ASIS 1988), vol. 25. American Society for Information Science, Atlanta (October 1988)

[5] Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. In: Machine Learning, pp. 143–175 (2001)

[6] Diederich, J., Kindermann, J., Leopold, E., Paass, G.: Authorship attribution with support vector machines. Applied Intelligence 19(1-2), 109–123 (2003)

[7] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)

[8] Ferilli, S., Biba, M., Basile, T.M., Esposito, F.: Combining qualitative and quantitative keyword extraction methods with document layout analysis. In: Post-proceedings of the 5th Italian Research Conference on Digital Library Management Systems (IRCDL 2009), pp. 22–33 (2009)

[9] Ferilli, S., Biba, M., Di Mauro, N., Basile, T.M.A., Esposito, F.: Plugging taxonomic similarity in first-order logic horn clauses comparison. In: Serra, R., Cucchiara, R. (eds.) AI*IA 2009. LNCS, vol. 5883, pp. 131–140. Springer, Heidelberg (2009)

[10] Ferilli, S., Leuzzi, F., Rotella, F.: Cooperating techniques for extracting conceptual taxonomies from text. In: Proceedings of the Workshop on MCP at AI*IA XIIth Conference (2011)

[11] Hamming, R.W.: Error detecting and error correcting codes. Bell System Technical Journal 29(2), 147–160 (1950)

[12] Jay, R., Jay, A.: Effective Presentation: How to Create and Deliver a Winning Presentation. Prentice Hall (2004)

[13] Jones, W.P., Furnas, G.W.: Pictures of relevance: A geometric analysis of similarity measures. Journal of the American Society for Information Science 38(6), 420–442 (1987)

[14] Karypis, G., Han, E.-H.(S.): Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Technical report, IN CIKM 2000 (2000)

[15] Klein, D., Manning, C.D.: Fast exact inference with a factored model for natural language parsing. In: Advances in Neural Information Processing Systems, vol. 15. MIT Press (2003)

[16] Leuzzi, F., Ferilli, S., Rotella, F.: Improving robustness and flexibility of concept taxonomy learning from text. In: Appice, A., Ceci, M., Loglisci, C., Manco, G., Masciari, E., Ras, Z.W. (eds.) NFMCP 2012 Workshop. LNCS (LNAI), vol. 7765, pp. 170–184. Springer, Heidelberg (2013)

[17] Maedche, A., Staab, S.: Mining ontologies from text. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 189–202. Springer, Heidelberg (2000)

[18] Maedche, A., Staab, S.: The text-to-onto ontology learning environment. In: ICCS-2000 - Eight ICCS, Software Demonstration (2000)
[19] Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13, 2004 (2003)
[20] Ogata, N.: A formal ontology discovery from web documents. In: Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.) WI 2001. LNCS (LNAI), vol. 2198, pp. 514–519. Springer, Heidelberg (2001)
[21] Cucchiarelli, A., Velardi, P., Navigli, R., Neri, F.: Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. In: Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press (2006)
[22] De Raedt, L., Kimmig, A., Toivonen, H.: Problog: a probabilistic prolog and its application in link discovery. In: Proceedings of 20th IJCAI, pp. 2468–2473. AAAI Press (2007)
[23] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at trec-3, pp. 109–126 (1996)
[24] Rotella, F., Ferilli, S., Leuzzi, F.: A domain based approach to information retrieval in digital libraries. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 129–140. Springer, Heidelberg (2013)
[25] Salton, G.: The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River (1971)
[26] Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill Book Company (1984)
[27] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18, 613–620 (1975)
[28] Sato, T.: A statistical learning method for logic programs with distribution semantics. In: Proceedings of the 12th ICLP 1995, pp. 715–729. MIT Press (1995)
[29] Singhal, A., Buckley, C., Mitra, M., Mitra, A.: Pivoted document length normalization, pp. 21–29. ACM Press (1996)
[30] Tweedie, F.J., Singh, S., Holmes, D.I.: Neural network applications in stylometry: The federalist papers. Computers and the Humanities 30(1), 1–10 (1996)
[31] Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on ACL, pp. 133–138. ACL, Morristown (1994)

Exploiting Wikipedia for Evaluating Semantic Relatedness Mechanisms

Felice Ferrara and Carlo Tasso

Artificial Intelligence Lab,
Department of Mathematics and Computer Science,
University of Udine, Italy
{felice.ferrara,carlo.tasso}@uniud.it

Abstract. The semantic relatedness between two concepts is a measure that quantifies the extent to which two concepts are semantically related. In the area of digital libraries, several mechanisms based on semantic relatedness methods have been proposed. Visualization interfaces, information extraction mechanisms, and classification approaches are just some examples of mechanisms where semantic relatedness methods can play a significant role and were successfully integrated. Due to the growing interest of researchers in areas like Digital Libraries, Semantic Web, Information Retrieval, and NLP, various approaches have been proposed for automatically computing the semantic relatedness. However, despite the growing number of proposed approaches, there are still significant criticalities in evaluating the results returned by different methods. The limitations evaluation mechanisms prevent an effective evaluation and several works in the literature emphasize that the exploited approaches are rather inconsistent. In order to overcome this limitation, we propose a new evaluation methodology where people provide feedback about the semantic relatedness between concepts explicitly defined in digital encyclopedias. In this paper, we specifically exploit Wikipedia for generating a reliable dataset.

1 Introduction

The terms *semantic similarity* and *semantic relatedness* (on which we focus in this paper) have often been used as synonyms in the areas of Natural Language Processing, Information Retrieval and Semantic Web, but some researchers highlighted significant differences between these two concepts. The concept of semantic relatedness is defined in the literature as the extent to which two concepts are related by semantic relations [17]. On the other hand, a possible definition of semantic similarity describes it as the measure which quantifies the extent to which two concepts can be used in an interchangeable way. According to this definition two semantically similar entities are also semantically related, but two semantically related concepts may be semantically dissimilar [3]. For example, the concepts of *bank* and *trust-company* are semantically similar and their similarity implies that they are also semantically related, but two concepts related by an

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 105–117, 2014.

antonymic[1] relation (such as the adjectives *bad* and *good*) are semantically related and semantically dissimilar. According to [20], semantic similarity is a more strict relation since it takes into account a focused set of semantic relations which are often stored in lexical ontologies such as Wordnet. In Wordnet, for example, synonyms[2] are grouped in synsets and a hierarchical structure connects hyponyms and hypernyms[3]. On the other hand, the semantic relatedness between two concepts depends on all the possible relations involving them. For example, in order to compute the semantic relatedness between two Wordnet concepts, we should use all the available semantic connections by including, for example, meronomy[4] and antonymy. However, two concepts can be related by more complex semantic relations which are usually not explicitly stored in lexical ontologies. Think, for example, to the case of two concepts that are semantically related by means of a chain of more than one semantic relation, involving other 'intermediate' concepts. For example, the pair *pope* and *Italy* can be related through the chain *pope* $\rightarrow$ *Vatican City* $\rightarrow$ *Rome* $\rightarrow$ *Italy*. This kind of relations is not explicitly included in Wordnet as well as all the other possible relations which can be entailed between concepts which are not directly related by standard relations. Moreover, it has to be noticed that humans organize their knowledge according to complex schemas by connecting concepts according to their background knowledge and experience [8]. The reasoning task where units of meaning are processed by the human mind in order to identify connections between concepts is referred in the literature as *evocation* [2], which can be also defined as the degree to which a concept brings to mind another one. Evocation adds cross-part-of-speech links among nouns, verbs, and adjectives [14]. Since the human mind works under the influence of personal experience, the evocation process builds relations which may be not true in an absolute way (for instance the relations between emotions and objects/animals) and this is why these relations cannot be available in knowledge bases such as Wordnet.

Obviously, all these aspects must be considered when we have to plan the evaluation of methods aimed at automatically quantifying the semantic relatedness (*SR methods*) or the semantic similarity (*SS methods*). Thesaurus-like resources , such as the Roget dataset [1], can be effectively used for evaluating the precision of SS methods: they connect terms by TR (Related-Term) links and by UF (Used-For) links, however such links are just a few for each term, whereas many others could be entailed.

For the above reasons, the feedback provided by humans about the relatedness between pairs of terms is commonly used in order to evaluate the precision of SR methods. However, the methodology currently used for both collecting feedback and evaluating precision of SR methods is widely criticized, even by the same

[1] Antonymy is the semantic relation which connects concepts with an opposite meanings.
[2] Two terms are synonyms if they have the identical or very similar meaning.
[3] A hyponym shares a *type of* relationship with its hypernym.
[4] The meronomy denotes a *part of* relation.

researchers who use it to analyze their results. These limitations are addressed in this work and, more specifically, the paper has two goals:

- describing the limitations of the state of the art mechanisms. A survey of the limitations of the approaches utilized for evaluating the accuracy of SR methods is given.
- proposing a new evaluation approach. We propose a new procedure aimed at effectively evaluating the precision of SR methods which analyze the content of Wikipedia, one of the main examples of Digital Library 2.0.

The choice of focusing on this specific digital library is mainly due to the growing interest of the research community on the usage of Wikipedia as knowledge source for computing semantic relatedness. In fact, the large coverage of concepts and the support to multilinguism makes Wikipedia very attractive for developing SR methods. Moreover, other researches point out that the refinements of the Wikipedia articles do not significantly influences the results of SR methods [19] while new concepts can be easily introduced and connected to the existing ones.

The paper is organized as follows: in Section 2 we describe the state of the art, major drawbacks are illustrated in mechanisms used for evaluating the precision of SR methods while the drawbacks of these approaches are the object of Section 3; in Section 4 we propose a new approach for facing these limitations; final considerations conclude the paper in Section 5.

2 Evaluating SR Methods: State of the Art

As reported in [3], three main approaches have been proposed in the literature for evaluating the precision of SR methods.

The approach utilized in [12] evaluates SR methods according to a set of qualitative heuristics. The simplest heuristic takes into account if the evaluated measure is a metric; in [9] the authors report a list of other suitable features for SR methods such as domain independence, independence from specific languages, coverage of included words, and coverage of the meanings of each word. The heuristic-based strategy is the simplest one but it also does not provide very significant results since it cannot numerically quantify the accuracy of results. For this reason, even if this strategy is a useful tool for designing new SR methods, it is not an effective tool for comparison [3].

More concrete results can be obtained by embedding SR methods in other hosting systems such as text clustering systems [11], metonymy resolution mechanisms [10], and recommender systems [5]. In these cases, different SR methods are compared and evaluated according to the improvement produced by the integration of the specific SR method within a larger system. However, it is quite clear that this strategy increases the difficulty in performing an extensive comparison of SR methods since: (i) different works face different tasks and use different datasets so preventing the repeatability of experiments and (ii) the computed precision can be influenced by the other components in the embedding system.

In order to overcome these drawbacks, a more direct strategy can be implemented by comparing the feedback of a set of humans with the results produced by SR approaches. The feedback of volunteers has been collected in order to create datasets which have been used in the majority of the works where the precision of SR methods has been evaluated. The first experiments aimed at creating this kind of datasets was exploited by Rubenstein and Goodenough [16]. In their experiments they exploited a deck of 65 cards where on each card there was a pair of nouns written in English. The researchers asked to 51 judges both to order the 65 pairs of words (from the most related pair to the most unrelated one) and to assign a score in [0.0,4.0] for quantifying the relatedness of each pair of terms. This experiment was also replicated by other researchers in different settings. One of the most popular dataset is the Related353 dataset [6] which is constituted by 353 word pairs is annotated with an integer in $[0, 10]$ by two sets of evaluators (composed by 13 and 16 judges respectively). Other works focused on the task of defining similar datasets for specific domains: in the biomedical field, Pedersen et al. collected the feedback of medics and physicians in order to evaluate SR methods in that specific domain [15]. Other works focused on generating larger datasets in an automatic way: in [18], a corpus of document is analyzed in order to extract pairs of semantically related terms by following the idea that pairs of terms which appear frequently in the same document are probably semantically related.

The numeric scores acquired in these experiments have been extensively used for evaluating the precision of SR methods. In order to reach this aim the Pearson product-moment and the Spearman rank order correlation coefficients have been used. The Pearson product-moment is a statistical tool used to check if the results of a SR method resemble human judgments. On the other hand, the comparison of two rankings of the pairs (the ranking which order the pairs according to the feedback provided by humans and the ranking which order the pairs according to the result of a SR method) can be executed by the Spearman coefficient. Both these coefficients have a numerical value in $[-1, +1]$, where -1 corresponds to completely uncorrelated rankings (low precision) and, conversely, $+1$ corresponds to a perfect correlation (high precision).

3 Drawbacks of the State of the Art

The experiments proposed in the literature mainly use datasets constituted by pairs of terms annotated by a group of humans. However, this approach has many criticalities which are emphasized even by the same researchers who adopted it. In this section we illustrate these limitations by organizing the discussion in two parts: in Section 3.1, we focus on the characteristics of the collections of pairs of terms and, in Section 3.2, we describe the features of both the human feedback and the procedures exploited for computing the precision of SR methods.

3.1 Characteristics of the Pairs of Terms

The quality of the feedback collected in the experiments referred in Section 2 strongly depends on the task submitted to the volunteers. The following points summarize the main limitations:

- **Shortage**. The dataset proposed by Rubenstein and Goodenough is constituted by only 65 pairs of nouns which cannot be used to exploit an extensive analysis for generalizing the findings. This limitation is partially faced by the Related353 dataset which is constituted by 353 pairs.
- **Terms instead of concepts**. The datasets are build up by terms which do not identify concepts. On the other hand, SR methods compute the semantic relatedness among concepts such as the synsets of Wordnet or the pages of Wikipedia. The proliferation of senses in knowledge bases such as Wordnet and Wikipedia makes hard the task of manually associating a sense to each term included in a dataset [17]. Consider, for example, that the term *love* is associated to 6 synsets in Wordnet and, on the other hand, in Wikipedia the term *love* identifies several senses: an emotion as well as people, songs, fictional characters, and movies. For tackling this problem, it is possible to manually associate some of the terms of the considered dataset to the Wikipedia concept that, most probably, was considered by the evaluators. On the other hand, in order to avoid the need for manual disambiguation of terms, the semantic relatedness between all the possible senses of the two terms can be identified and fixed in the following way: the pair of senses with the highest semantic relatedness computed by the evaluated SR method is considered for assigning two specific senses to the two terms. Both these approaches are questionable since the judges were not conscious of all the various meanings of the words when they annotated the pairs.
- **Uncovered domains and semantic relations**. The datasets created by Rubenstein and Goodenough as well as the Related353 dataset were defined with the main goal of covering many possible degrees of similarity. Following this idea, the authors used very general terms without taking into account the idea of choosing terms in different domains. This is limitation which prevents the generalization of the results. In particular, we highlight that the information provided or extracted from a knowledge base may differ according to the given topic. We can imagine that in Wikipedia, for example, some topics are described better than others. It is also possible that different knowledge bases (such as Wikipedia, Wordnet or other ontologies) may provide better results in different domains. For this reason it would be interesting to have datasets where pairs of terms are associated to domains or at least to have datasets where several distinct domains are covered. Similarly, a more reliable approach should also take care of covering a sufficient set of semantic relations. In fact specific SR method could be adequate for catching a specific semantic relation but it could not work with other relations. This information is obviously missing also in datasets created in an automatic way.

3.2 Characteristics of the Feedback and Evaluation Procedure

The agreement among the evaluators is used in the literature for estimating the quality of the collected feedback: this follows the idea that higher is the agreement more reliable is the collected feedback. According to the literature, the level of agreement is sufficient to assess the precision of SR methods. However, there is not a threshold for the required agreement between the judges and this is also true for domain-dependent datasets. Moreover, also other features of the feedback collected from humans may greatly influence negatively the quality of the evaluation. More specifically, we identify the following points:

- **Pairs with low agreement.** Different works use different strategies to manage pairs ot terms with low agreement among judges. An example of these pairs is *(monk,oracle)* in the Related353 dataset which was annotated by 13 evaluators who returned the following votes $(7, 8, 3, 4, 4, 6, 5, 8, 6, 3, 4, 6, 1)$. In the majority of the works available in the literature these pairs are threaten exactly like the others, but in [15] the authors proposed to discard pairs with a very low agreement in order to have more significant results. Obviously, this idea can be applied only when the dataset is constituted by a large set of pairs. This is a very important issue since, as noticed in [3], the available datasets show a significant agreement only when the existence of the semantic relation is very clear (for instance the terms are synonyms or they are completely unrelated).
- **The choice of the scale.** The choice of the scale for collecting the feedback is a controversial point and has a strong impact on the agreement among the judges. By adopting a very fine-grained scale the judges have many possible choices and they can provide more accurate responses. This was the motivation for the approach proposed by Rubenstein and Goodenough who also asked people to order the pairs in order to have more coherent responses. In fact, by ordering the pairs each judge could assign a decreasing list of values to quantify the semantic relatedness. However, this mechanism does not scale up to a large set of pairs since it requires a huge workload for ordering many pairs of terms. For this reason, in the task for acquiring the feedback for larger datasets like the Related353 dataset it is not asked to the evaluators to order the pairs. In this case, the humans could not rely on the order imposed to the pairs for assigning a vote and, consequently, it was harder for them to be coherent with previously assigned votes. For this reason, when the judges only annotate pairs of terms with a number it is better to avoid very fine-grained scale in order to have more consistent responses.
- **Bias introduced by specific communities.** Different communities of evaluators may evaluate the semantic relatedness between two concepts according to different perspectives. This is clearly reported in [15] where the authors show that physicians and medics judged differently the semantic relatedness between terms in the field of biology. On the other hand, it makes sense to evaluate SR methods only on pairs where the feedback is not biased by the perspective of a specific community.

– **Metric robustness.** The Pearson coefficient is a statistical tool used to catch the strength of the linear correlation between the human judgments and the score computed by a specific SR method. However, the correlation between the votes provided by humans and the SR method can be nonlinear. Moreover, the Pearson correlation is based on the assumption that the two compared random variables are normally distributed, whereas the actual distribution of the relatedness values is at the moment unknown [3]. On the other hand, the Spearman coefficient, which does not directly compare human votes with the results of the SR method, seems to be more robust.

4 Toward a New Evaluation Strategy

In order to face the limitations described in the previous section we propose a new strategy for evaluating SR methods. In this section we describe our ongoing work (Section 4.1) as well as our future steps (Section 4.2).

4.1 New Resources and Procedures

As already mentioned, other researchers showed that humans can judge the semantic relatedness by using a numerical estimation only if the answer is quite obvious. In fact, the experiments described in Section 2 showed that the agreement among the judges was significantly hight only when the pairs were composed by two synonyms or by two completely unrelated terms. Our hypothesis is that humans can perceive the semantic relatedness, but they are not used to quantify it by using a number. The difficulty in acquiring reliable feedback from humans is mainly due to the problem of having datasets constituted by terms which may be polysemic, i.e. having multiple senses. Starting from this assumption, here we propose a new procedure for collecting more significant responses from the judges, by avoiding both expensive workload, such as ordering a long sequence of pairs of terms, and tricky/noisy tasks, such as selecting a numeric level to quantify the semantic relatedness among two terms.

Our proposal is to ask judges to select the concept (from a set of proposed concepts) which is most related to a given concept, where each concept is associated to a specific knowledge base (in our current work concepts are identified by Wikipedia pages). By associating each term to a concept of a knowledge base we can overcome the limitation of having datasets constituted by only terms. This approach allows to obtain two advantages: (i) the judges can take into account a unique specific meaning of the concepts when they produce their responses and (ii) the evaluated SR method can exploit the Wikipedia page associated to the concept for computing the semantic relatedness.

More technically, we defined the questions for the judges as triples $T = (t_1, \ldots, t_m)$, where the triple $t_i = (target_i, c_{i1}, c_{i2})$ is constituted by a target concept and two other concepts c_{i1}, c_{i2}. For each triple t_i, the judges have to identify which one among c_{i1} and c_{i2} is (in their views) more related to $target_i$. For example, given the triple t=(*Musician*, *Watch*, *Trumpet*), the evaluator can

select *Watch* or *Trumpet* as more related to *Musician*. The reader can notice that the proposed procedure does not depend on a specific scale for collecting the feedback and this also simplifies the work of the judges who have to select only the most related concept. By selecting the concept semantically more related to the target concept the judge orders the three concepts according to the relatedness to the target. By following the previous example, if a judge chooses *Trumpet* then he implicitly defines the ordered list of concepts (*Musician*, *Trumpet*, *Watch*) since *Trumpet* has been considered as more related to *Musician* than *Watch*. We then take into account the way the judges order the concepts in each triple for evaluating the precision of a SR method. In particular, the SR method can order the concepts in the triple $t_i = (target_i, c_{i1}, c_{i2})$ by computing the semantic relatedness between the target concept and the two concepts c_{i1} and c_{i2} and, consequently, it can produce a rank. In this way we compute the precision of the SR method according to the percentage of cases in which the SR method orders the concepts of the triples as the humans did.

However, we also believe that there are various cases where humans cannot provide a response: the judge may be not familiar with a concept or even a topic or two concepts may be (more or less) equally semantically related to the target concept. In order to manage these situations, the judges are allowed to skip the evaluation of a triple, since we are keen to identify the responses for which the judges are sufficiently confident. By taking into account the number of judges who skipped a triple, we can measure a degree of trustworthiness of the overall feedback acquired for a specific single triple. More specifically, for each triple we computed an *Indecision Score* as the ratio between the number of judges who skipped the triple and the total number of judges. By taking into account a maximum threshold on the Indecision Score, we then remove the triples for which there is a certain percentage of judges who did not provide a response. We also filter out the triples with a low agreement among judges, by following the idea that a low agreement can be the result of different evaluation perspectives. To this aim we computed, for each triple t_i, an *Agreement Score* as the maximum between (i) the ratio between the number of judges who selected the concept c_{i1} and the total number of judges and (ii) the ratio between the number of judges who selected the concept c_{i2} and the total number of judges. By requiring an Agreement Score higher than a certain threshold, we can remove ambiguities which may be introduced by different communities with different perspectives.

We identify for each triple a 'correct' order of the concepts by taking into account the order defined by the majority of the judges. For example, supposing that the concept *Trumpet*, in the triple t=(*Musician*, *Watch*, *Trumpet*), is more frequently selected than the concept *Watch*, then the order (*Musician*, *Trumpet*, *Watch*) is taken as the correct ranking. This 'correct ranking' is compared to the order computed by the evaluated SR method. In this way we define the *precision* of the evaluated SR method as the percentage of the correctly ordered triples by the SR method.

Obviously, the approach used to build the triples has a significant impact on the results. As we said in Section 3, one of the main drawbacks of the datasets

described in the literature depends on the number of domains and of different semantic relations included in the dataset. In order to face this issue, we have defined a specific set of templates for the triples, such as ($\langle TARGET\rangle$, $\langle Emotion_1\rangle$, $\langle Emotion_2\rangle$) and ($\langle TARGET\rangle$, $\langle Work_1\rangle$, $\langle Work_2\rangle$). Then, we create some triples by creating instantiating each template. For example, from the template ($\langle TARGET\rangle$, $\langle Emotion_1\rangle$, $\langle Emotion_2\rangle$), we can build the triple (*Love*, *Graditude*, *Jelausy*), the triple (*Clown*, *Humor*, *Fear*) and so on. We also include other triples by picking concepts from systems such as Delicious and Open Directory. In particular, tags, categories, and other terms are extracted from these systems in order to create new triples. By using stacks of Delicious and categories in Open Directory we also select concepts (that must be concepts of Wikipedia) belonging to different domains. In this way we face (at least partially) the problem of covering semantic relations in different domains. In our first experiments we collected the feedback of 10 judges and each of them evaluated 420 triples in a month. Since the Agreement Score measures the agreement among the judges only on a single triple, we evaluated the overall agreement among the judges by by means of the Fleiss'kappa [7]. The Fleiss'kappa allows us to measure the agreement of the judges over the entire set of triples and, according to this analysis, we have a significant agreement also over the entire set of triples (kappa=0.783). Then we filtered the triples by throwing out the triples with an Agreement Score lower then 0.7 (i.e. we require that at least 7 of the 10 judges provided the same response) and with an Indecision Score higher then 0.2 (i.e. we require that at maximum 2 judges skipped the question). As expected, after this filtering step, we have that the agreement among the judges increases (kappa=0.849). However, it is interesting to observe that our filtering interventions removed only 27 triples from the initial set of 420 triples. Two examples of these triples are (*Mammal*, *Dolphin*, *Lion*) and (*Lifeguard*, *Holiday*, *Work*). These two triples show the usefulness of the filtering step since we observed that many judges skipped the triple (*Mammal*, *Dolphin*, *Lion*) since *Dolphin* and *Lion* are both *Mammals*. The Indecision Score is used to discard this triple because, in this case, people could not find semantic relations for identifying which one of the two concepts is more related to the target concept. Similarly, if two concepts are completely unrelated to a given target concept, then judges cannot find semantic relations for answering. The Indecision Score allows us to remove these triples avoiding in this way potential ambiguities. In the case of the triple (*Mammal*, *Dolphin*, *Lion*), a part of the judges considered *Lifeguard* as someone you can meet during *Holiday* whereas another part of the judges considered *Lifeguard* as a *Work*. In this case, there is a low agreement due to the subjective way of perceiving the semantic relatedness. The Agreement Score allows us to remove such triples, enhancing in this way the significance of the dataset.

4.2 Ongoing Evaluation and Future Steps

At the moment we are comparing new SR methods (that we specifically designed in order to compute the semantic relatedness between the Wikipedia concepts) with other state of the art mechanisms.

In particular we defined new SR methods by extending some approaches proposed [13] where: (i) each Wikipedia concept is represented by its incoming pages (i.e. the pages with a link to the concept) and outcoming pages (i.e. the pages linked by the concept) and (ii) the semantic relatedness among two pages is computed by comparing their corresponding representations (larger is the number of shared incoming/outcoming links, higher is the similarity among the concepts). In particular, in this work we propose two metrics which will be referred as CIN and GDOUT in the rest of the paper.

The CIN metric describes each concept as a weighted vector of Wikipedia pages. In particular given a concept of Wikipedia, the pages which have a link to the concept describe it and the weight of each page (i.e. each component of the vector) is equal to $log(\frac{|W|}{|T|})$ where W is the set of pages in Wikipedia and T is the number of articles linked by the specific page (i.e. the specific component of the vector). Given such representation of concepts, the CIN metric computes the semantic relatedness between two concepts as the cosine similarity between the two corresponding vectors. In this way, the metric computes the semantic relatedness among two concepts according to the shared incoming pages. The metric assumes that the concepts having many outcoming pages are less specific and, for this reason, the semantic relatedness among two concepts is higher when the corresponding Wikipedia pages share many incoming pages with few outcoming pages.

On the other hand, the GDOUT metric is based on a different assumption. In fact, in this case, the semantic relatedness between two concepts is estimated by taking into account the number of outcoming pages shared between the corresponding Wikipedia pages. More technically, the GDOUT metric uses a variation the Normalized Google Distance [4] for computing the semantic relatedness between the concepts a and b as

$$GDOUT = 1 - \frac{log\left(max\left(|A|, |B|\right)\right) - log\left(|A \cap B|\right)}{log\left(|W|\right) - log\left(min\left(|A|, |B|\right)\right)}$$

where A is the set of pages linked by the concept a, B is the set of pages linked by the concept b, and W is still the set of pages available in Wikipedia.

We utilized our approach in order to evaluate the results produced by these two SR methods and we obtained that the CIN approach has a higher precision (the precision is equal to 0.87) than the GDOUT method (the precision is equal to 0.80 in this case).

At the moment we are working on utilizing the datasets constituted by pairs of terms and on embedding the SR methods in different systems in order to compare our evaluation approach with other evaluation approaches. In particular, we are interested in embedding the SR methods also in other different systems in order to verify if the results changes according to the system where the SR methods are integrated.

In order to promote a more exhaustive evaluation campaign of the SR methods proposed in the literature we are also working on other two possible extensions of our proposal.

First, we are interested in collecting a larger set of responses by utilizing crowdsourcing systems such as Amazon Mechanical Turk. In particular, we are interested in evaluating if different levels of agreement among the judges can be found by utilizing a new, different, and larger set of judges. These future experiments will allow us to better evaluate also the statistical relevance of our current results.

Second, we recognize that Wikipedia is not the only possible knowledge source which can be used for computing the semantic relatedness. For example, other works in the literature compute the semantic relatedness among the synsets of Wordnet. In order to have an exhaustive evaluation campaign we need to have triples constituted by the concepts defined in other knowledges sources such as Wordnet. We are evaluating two possible strategies. The first one is to repeat our work for constructing a new dataset which cover concepts of Wordnet. On the other hand, we have designed and developed an intelligent framework which can support the alignment of the concepts of Wikipedia to the synsets in Wordnet. By exploiting this tool we aim at associating the concepts in our dataset to Wordnet synsets.

5 Conclusion

Many tools and approaches which integrate the computation of SR among concepts have been proposed in the literature in order to improve the access to digital libraries [21]. On the other hand, in this paper, we (i) analyzed the limitations of the approaches traditionally utilized to evaluate the precision of SR methods and (ii) proposed a new approach for producing more reliable datasets and evaluations. Our first results about the agreement among the judges and the pruning of ambiguous triple seem promising.

Future works will investigate the usage of crowdsourcing systems for collecting larger set of responses from a larger set of judges. We will also study the problem of using concepts available in knowledge sources different from Wikipedia by associating concepts of Wikipedia to concepts of Wordnet.

References

1. Roget's 21st century thesaurus, 3rd edn. (October 2012), `http://thesaurus.com/browse/dataset`
2. Boyd-graber, J., Fellbaum, C., Osherson, D., Schapire, R.: Adding dense, weighted connections to wordnet. In: Proceedings of the Third International WordNet Conference (2006)
3. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Comput. Linguist. 32(1), 13–47 (2006)
4. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Trans. on Knowl. and Data Eng. 19(3), 370–383 (2007)

5. Ferrara, F., Tasso, C.: Integrating semantic relatedness in a collaborative filtering system. In: Proceedings of the 19th Int. Workshop on Personalization and Recommendation on the Web and Beyond, pp. 75–82 (2012)
6. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. ACM Trans. Inf. Syst. 20(1), 116–131 (2002)
7. Fleiss, J.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI 2007, pp. 1606–1611. Morgan Kaufmann Publishers Inc., San Francisco (2007)
9. Gracia, J., Mena, E.: Web-based measure of semantic relatedness. In: Bailey, J., Maier, D., Schewe, K.-D., Thalheim, B., Wang, X.S. (eds.) WISE 2008. LNCS, vol. 5175, pp. 136–150. Springer, Heidelberg (2008)
10. Hayes, J., Veale, T., Seco, N.: Enriching wordnet via generative metonymy and creative polysemy. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, pp. 149–152. European Language Resources Association (2004)
11. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 389–396. ACM, New York (2009)
12. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, ACL 1998, vol. 2, pp. 768–774. Association for Computational Linguistics, Stroudsburg (1998)
13. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, pp. 25–30. AAAI Press (2008)
14. Nikolova, S., Boyd-Graber, J., Fellbaum, C.: Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools. In: Mehler, A., Kühnberger, K.-U., Lobin, H., Lüngen, H., Storrer, A., Witt, A. (eds.) Modeling, Learning, and Proc. of Text-Tech. Data Struct. SCI, vol. 370, pp. 81–93. Springer, Heidelberg (2011)
15. Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics 40(3), 288–299 (2007)
16. Rubenstein, H., Goodenough, J.B.: Contextual correlates of synonymy. Commun. ACM 8(10) (October 1965)
17. Strube, M., Ponzetto, S.P.: Wikirelate! computing semantic relatedness using wikipedia. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 2, pp. 1419–1424. AAAI Press (2006)
18. Zesch, T., Gurevych, I.: Automatically creating datasets for measures of semantic relatedness. In: Proceedings of the Workshop on Linguistic Distances, LD 2006, pp. 16–24. Association for Computational Linguistics, Stroudsburg (2006)

19. Zesch, T., Gurevych, I.: The more the better? assessing the influence of wikipedia's growth on semantic relatedness measures. In: Calzolari, N. (ed.) Proceedings of the Seventh International Conference on Language Resources and Evaluation. European Language Resources Association, Valletta (May 2010)
20. Zesch, T., Gurevych, I.: Wisdom of crowds versus wisdom of linguists; measuring the semantic relatedness of words. Nat. Lang. Eng. 16(1), 25–59 (2010)
21. Zhang, W., Feng, W., Wang, J.: Integrating semantic relatedness and words' intrinsic features for keyword extraction. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 2225–2231. AAAI Press (2013)

Semantic Lenses as Exploration Method for Scholarly Articles

Silvio Peroni[1], Francesca Tomasi[2], Fabio Vitali[1], and Jacopo Zingoni[1]

[1] Department of Computer Science and Engineering, University of Bologna, Italy
{essepuntato,fabio}@cs.unibo.it, jacopo.zingoni@gmail.com
[2] Department of Classical Philology and Italian Studies, University of Bologna, Italy
francesca.tomasi@unibo.it

Abstract. In a move towards an enrichment of the metadata models that are used in the electronic publication of scholarly literature, modern publishers are making steps towards *semantic publishing*. The possibility to explore a collection of scientific papers (a digital library, a repository or an archive of data) using different and multiple facets, i.e., different and multiple points of view on the digital collection, increases on the one hand the success of information retrieval and on the other hand the availability of richer data sets. Multiple facets are the natural navigation method made possible by an adequate ontological representation of a class of homogeneous documents. Context and content of published journal articles are thus components that in the representation of information at the metadata level constitute a fundamental approach to semantic enhancement. In this paper we introduced a test in using a particular semantic publishing model, called *semantic lenses*, to semantically enhance published journal articles.

Keywords: context and content, document semantics, semantic publishing.

1 Introduction

It is a truism to assert that the richness of the metadata model used in digital collections is instrumental in expanding and enhancing the uses made possible by them on the collection, and that models that are too simple may well result in widespread adoption, but on the other hand provide a weak representation of the information contained in the collection, and may induce conceptual errors and misrepresentations, as we discussed (among many) in [1].

Nowhere this is more visible than in the *publishing domain*. Publishers started to use the Web as distribution channel since its early origins [2], but their market exploded with the advent of XML-based languages (e.g. (X)HTML and DocBook), ebook formats (e.g. EPUB and PDF), online vendors (e.g. Apple's iTunes bookstore and Amazon's bookstore), and tablet reading devices (e.g. iPad and Kindle). Similarly, metadata associated to electronic publications while inheriting the results of a multisecular discipline, library studies, have managed to

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 118–129, 2014.

coalesce into a number of very simple, minimal, models, such as Dublin Core [3], that although pretty successful are inevitably crippled by their own simplicity.

Simultaneously to the evolution of the Web into the Semantic Web, modern publishers (and in particular scholarly publishers) have taken steps to enhance their digital publications with semantics, an approach that is known as *semantic publishing* [4]. In brief, semantic publishing is the use of Web and Semantic Web technologies to represent formally the meaning of a published document, by specifying a large quantity of information about it as metadata and to publish them as Open Linked Data. As a confirmation of this trend, recently the Nature Publishing Group (publisher of *Nature*), the American Association for the Advancement of Science (publisher of *Science*) and the Oxford University Press have all announced initiatives to open their articles' reference lists and to publish them as Open Linked Data[1,2]. The open archive movement[3] is increasing in the field scientific papers publishing and big commercial companies, like i.e. Springer, support the idea of "institutional repositories" and the concept of "open access publishing" as a solution that "makes your work immediately and permanently available online for everyone, everywhere"[4]. A significant increase of open access journals reveal the impact of new methods of digital publishing[5].

Open archives as repositories for the dissemination, the interchange and the preservation of scholarly articles and related metadata but also open access as method of publication are becoming a strategy and a paradigm in the field of publishing. Digital libraries of scientific papers use these techniques, theories and methods in order non only to speed up the access given to publications but also to increase the amount of digital data i.e. research articles, they can associate to such articles. Even editors and publishers that did not marry into the open access philosophy are creating digital collections of scientific papers under the guise of freemium platforms for accessing for free the metadata related to their publication, and pay for the full-text of the articles. Even many aggregation platforms (e.g. Elsevier Science and Emerald[6]) found in the possibility to give access to big collections of scientific papers a new way of exploring knowledge.

Many of these platforms are defining semantic models to enhance the digital representation of their articles. However, this enhancement is not a straightforward operation, since it involves much more that simply making semantically

[1] Science joins Nature in opening reference lists: `http://opencitations.wordpress.com/2012/06/16/science-joins-nature-in-opening-reference-citations`

[2] Oxford University Press to support Open Citations: `http://opencitations.wordpress.com/2012/06/22/oxford-university-press-to-support-open-citations`

[3] `http://www.openarchives.org`

[4] `http://www.springer.com/open+access?SGWID=0-169302-0-0-0`

[5] See the Directory of Open Access Journals: `http://www.doaj.org/doaj?func=newTitles&fromDate=2012-11-14+00%3A00%3A00&untilDate=2012-12-14+19%3A15%3A31`

[6] Elsevier Science (`http://www.sciencedirect.com`), Emerald (`http://www.emeraldinsight.com`).

precise statements about named entities within the text. For instance, the sentence "Christopher Marlowe was the real author of many Shakespeare's plays" has possibly one formal representation, but its use in a scholarly document may be characterized in many different ways, as a claim, an hypothesis, a rebutted concept, or even as an example of an English sentence in a paper not discussing Shakespeare's plays at all (as this one).

In [5], we showed how several relevant interpretation layers exist beyond the bare words of a scientific paper – such as the context of the publication, its structural components, its rhetorical structures (e.g. Introduction, Results, Discussion), or the network of citations that connects the publication to its wider context of scholarly works. These points of view are usually combined together to create an effective unit of scholarly communication so well integrated into the paper as a whole and into the rhetorical flow of the natural language of the text, so as to be scarcely discernible as separate entities by the reader.

In this paper we use a well-known scholarly paper, *DelosDLMS – The Integrated DELOS Digital Library Management System* by Agosti *et al.* [6], to investigate the feasibility and the usefulness of separating these aspects into eight different sets of machine-readable semantic assertions (called *semantic lenses*), where each set describes one of them, from the most contextual to the most document-specific: research context, authors' contributions and roles, publication context, document structure, rhetoric organization of discourse, citation network, argumentative characterisation of text, and textual semantics.

The rest of the paper is organised as follows. In Section 2 we introduce some significant works related to semantic publishing experiences and models. In Section 3 we show an application of semantic lenses onto a particular scholarly article. Finally (Section 4) we conclude the paper sketching out some future works and briefly present a prototype named TAL (*Through A Lens*), an HTML interface for scholarly papers.

2 Related Works

Much current literature concerns both the proofs of concepts for semantic publishing applications and the models for the description of digital publishing from different perspective. Because of this richness, here we present just some of the most important and significant works on these topics.

In [4], Shotton *et al.* describe their experience in enriching and providing appropriate Web interfaces for scholarly papers enhanced with provenance informations, scientific data, bibliographic references, interactive maps and tables, with the intention to highlights the advantages of semantic publishing to a broader audience. Along the same lines, in their work [7] Pettifer *et al.* introduce pros and cons of the various formats for the publication of scholarly articles and propose an application for the semantic enhancement of PDF documents according to established ontologies.

A number of vocabularies for the description of research projects and related entities have been developed, e.g. the VIVO Ontology[7] – researched for

[7] VIVO Ontology: `http://vivoweb.org/ontology/core`

describing the social networks of academics, their research and teaching activities, their expertise, and their relationships to information resources – and DOAP, the *Description Of A Project*[8] – an ontology with multi-lingual definitions that contains terms specific for software development projects.

One of the most widely used ontology for describing bibliographic entities and their aggregations is BIBO, the *Bibliographic Ontology* [8]. FRBR, *Functional Requirements for Bibliographic Records* [9], is yet another more structured model for describing documents and their evolution in time. One of the most important aspects of FRBR is the fact that it is not tied to a particular metadata schema or implementation.

Several works have been proposed in the past to model the rhetoric and argumentation of papers. For instance, the SALT application [10] permits someone such as the author "to enrich the document with formal descriptions of claims, supports and rhetorical relation as part of their writing process". There are other works, based on [11], that offer an application of Toulmin's model within specific scholarly domains, for instance the legal and legislative domain [12]. A good review of all the others Semantic Web models for the description of arguments can be found in [13]. A comprehensive analysis of the application of Semantic Web ideas and techniques in digital repositories can be read in [14].

3 Context and Content through Semantic Lenses

In [5] we introduced the idea that the semantics of a document is definable from different perspectives, where each perspective is represented as a *semantic lens* that is *applied* to a document to reveal a particular semantic facet.

A faceted classification system [15] in the field of library science is a bottom-up scheme that divides a subject into concepts and gives rules to use these concepts in constructing a structured subject. This approach makes it possible use a kind of poly-hierarchical relationship between the elements of the description [16].

But facets have to be transformed in an ontology in order to give access to the deep meaning of the documents. An ontology has been defined[9] to formally define these lenses so as to allow the annotation of resources such as scholarly papers. In addition, since the application of the semantic lenses to a document is an *authorial activity*, i.e. the action of a person (the original author as well as anyone else) taking responsibility for a semantic interpretation of the document, we also need to record theprovenance of the semantic statements according to the *PROV Ontology* (*PROV-O*) [17].

In the following subsections we introduce the lenses using the well-known paper *DelosDLMS – The Integrated DELOS Digital Library Management System* by Agosti et al. [6] as the scholarly article on which the small snippets of semantic lenses are based.

[8] DOAP: `http://usefulinc.com/ns/doap`
[9] Lens Application Ontology (LAO): `http://www.essepuntato.it/2011/03/lens`

3.1 Describing the Context

Writing a scientific paper is usually the final stage of an often complex collaborative and multi-domain activity of undertaking the research investigation from which the paper arises. The organizations involved, the people affiliated to these organizations and their roles and contributions, the grants provided by funding agencies, the research projects funded by such grants, the social context in which a scientific paper is written, the venue within which a paper appears: all these provide the research *context* that leads, directly or indirectly, to the genesis of the paper, and awareness of these may have a strong impact on the credibility and authoritativeness of its scientific content.

The concept of context is a polysemic textual situation because it runs across a variety of different disciplines. In general "the broad notion of context [is] constituted by the interactions and relationships between a TE [target entity] and its environment" [18]. In particular in the archival domain this concept regards the need to separate the description of document from the description of people that create the documents. The EAC-CPF (Encoded Archival Context-Corporate Bodies, Persons and Families) is a DTD[10], an XML Schema[11] and now an ontology [19] for translate the ISAAR (CPF), the International Standard for Archival Authority Records [20], in a formal language.

Daniel Pitti states that "relations between records, creators, and functions and activities are dynamic and complex, and not fixed and simple. Creators are related to other creators. Records are related to other records. Functions and activities are related to other functions and activities. And each of these is interrelated with the others. [...]. By developing dedicated semantics and structures for describing each descriptive component and its complex interrelations, we can build descriptive systems that are far more efficient and effective than those we have realized in print" [21]. So the context reflects the need to separate the object (the paper) from the information surrounding it, and in fact the context reflects the relationships between data and structured metadata, but is also an interpretation key of the document as a complex entity whose information emerges only when analysing the elements of the document in their specific context.

Given these assumptions, we need to point out that semantic lenses have to be used as a complex system, in a network perspective of interconnected scopes, rather than as a hierarchical model of independent layers.

Three lenses are designed to cover the contextual aspects of a scholarly text:

- *Research context*: the background from which the paper emerged (the research described, the institutions involved, the sources of funding, etc.).
- *Contributions and roles*: the individuals claiming authorship on the paper and what specific contributions each of them provided.
- *Publication context*: any information about the event (e.g. conference or workshop) and publication venue of the paper (such as the proceedings or the journal), as well as connections to the other papers sharing the same event or venue.

[10] ISAAR(CPF) DTD: `http://www3.iath.virginia.edu/eac/`
[11] ISAAR(CPF) Schema: `http://eac.staatsbibliothek-berlin.de`

Using [6] as the basis for the annotations example, we describe the *contextual environment*, that is the *research context,* that made possible writing this paper by using[12] *FRAPO*, the *Funding, Research Administration and Projects Ontology*[13], as shown in the following excerpt. The excerpt specifies that the European Commission, as a funding agency *(#1)*, funded the network of excellence DELOS (*#2*) that led to the aforementioned paper (*#3*)[14]:

```
:research-context {
:ec a frapo:FundingAgency ; foaf:name "EU Commission" ; #1
  frapo:funds [ a frapo:Endeavour ; #2
    foaf:name "A Network of Excellence on Digital Libraries";
    frapo:hasOutput :delosdlms ] . } #3
```

Then we use *SCoRO* (the *Scholarly Contributions and Roles Ontology*[15]) and its imported ontology *PRO* (the *Publishing Roles Ontology*[16]) [22] to identify the *roles and contributions.* Once again, in order to be concise, only the code for one of the many contributors will be shown, in this case for the first one, Maristella Agosti. We can identify her *role* (e.g. being affiliate with the University of Padua during the realization of the paper – *#4*) and her *contribution* (*#5*) within the context of this paper.

```
:contributions-and-roles {
:agosti a foaf:Person ; foaf:name "Maristella Agosti" ;
  pro:holdsRoleInTime [ a scoro:OrganizationalRole ; #4
    pro:withRole scoro:affiliate ;
    pro:relatesToOrganization [ a frapo:University ;
      foaf:name "University of Padua" ] ;
    pro:relatesToDocument :delosdlms ],
  scoro:makesContribution [a scoro:ContributionSituation; #5
    scoro:withContribution scoro:writes-manuscript-draft ;
    scoro:withContributionEffort scoro:major-effort ;
    scoro:relatesToEntity :delosdlms ] }
```

We then describe the *publication context* of the paper using FaBiO, the *FRBR-aligned Bibliographic Ontology* [23] and BiRO, the *Bibliographic Reference Ontology*[17], specifying the conference proceedings in which the paper was published

[12] Note that all the ontologies used or suggested in this paper to describe "lenses" statements have been chosen as an appropriate and convincing example of an ontology that fulfils the requirements for the lens, since they allow us to fully describe all the document aspects we are interested in. However, their choice is not unique, and many other ontologies may exist to fulfil the same role, so as to allow the use of other models (such as those described in Section 2) instead of them.

[13] FRAPO: `http://purl.org/cerif/frapo`

[14] This and the following RDF examples are written in Turtle (`http://www.w3.org/TeamSubmission/turtle/`), with namespace definitions defined at `http://www.essepuntato.it/2013/tal/prefixes`.

[15] SCoRO: `http://purl.org/spar/scoro`

[16] PRO: `http://purl.org/spar/pro`

[17] FaBiO: `http://purl.org/spar/fabio`; BiRO: `http://purl.org/spar/biro`.

(*#6*) and the list of its references to other related documents *(#7)* – which is crucial for semantic publishing:

```
:publication-context {
# The textual realisation of the paper
:version-of-record a fabio:ConferencePaper ; #6
  frbr:realisationOf :delosdlms ;
  dcterms:title "DelosDLMS - The Integrated DELOS Digital
      Library Management System" ;
  prism:doi "10.1007/978-3-540-77088-6_4" ;
  frbr:partOf [ a fabio:ConferenceProceedings ;
    dcterms:title "Proc. 1st International DELOS Conference";
    fabio:hasPublicationYear "2007"^^xsd:gYear ]
  frbr:part [ a biro:ReferenceList ; #7
    co:element [ biro:references
      <http://dx.doi.org/10.1109/ICCV.1998.710779> ] ... ]}
```

3.2 Describing the Content

The semantics of *the content* of a document, i.e. the semantics that is implicitly defined in and inferable from the text, can be described from different points of view. For example, the overall *structure* of the text – i.e. the organization of the text of the document into structured containers, blocks of text, inline elements – is often expressed by means of markup languages such as XML and LaTeX, that have constructs for describing content hierarchically.

In the field of textual editing, the TEI schema [24] represents a standard model for the encoding of humanistic texts using an embedded markup. The *Guidelines* elaborated in the TEI project reflect on different aspects of the interpretative intervention of the editor in describing textual entities. A big effort is now devoted towards the translation of this XML Schema into an ontology in the domain of cultural heritage, mapping TEI onto CIDOC-CRM [25] a conceptual model for describing entities used in cultural heritage documentation.

In an Semantic Web context, we would rather use an ontology that describes the markup structures in OWL. For this we need a way to separate the document from its interpretation, i.e., a way to apply a meta-syntax for stand-off annotations of textual content with fully W3C-compliant technologies. For this reason, we use *EARMARK* [26], an ontology[18] of a markup metalanguage, to describe the structure of the document as a set of OWL assertions to associate formal and explicit semantics [27]. Through the *Pattern Ontology* (*PO*)[19] [28] in combination with EARMARK we can associate a particular structural semantics to markup elements, such an element *h3 expressing* the concept of being a block of text (*#9*), or the *div* element containing it being a container with an header (*#8*), as shown in the following:

[18] EARMARK: `http://www.essepuntato.it/2008/12/earmark`
[19] PO: `http://www.essepuntato.it/2008/12/pattern`

```
:structure { :div a earmark:Element ;
  la:expresses pattern:HeadedContainer ; #8
  earmark:hasGeneralIdentifier "div" ;
  c:firstItem [ c:itemContent ... ; c:nextItem [
  c:itemContent :h-sec-2 ; ... c:nextItem [ ...
  c:itemContent :p4-sec-2 ... ] ] ] .
:h-sec-3-1 a earmark:Element #9
  la:expresses pattern:Block ;
  earmark:hasGeneralIdentifier "h3" ;
  c:firstItem [ c:itemContent :r-h-sec-3-1 ] .
# Text node within :h-sec-3
:r-h-sec-3-1 a earmark:PointerRange ...
:p1-sec-3-1 a earmark:Element # Sec 3.1, Par 1
  la:expresses pattern:Block ;
  earmark:hasGeneralIdentifier "p" ... }
```

Just a little above a purely structural perspective, we place the identification and organization of the *rhetorical components* of the text, such as a section being an *Introduction*, some paragraphs describing the *Methods* of the research, or the presented *Results* or the paper's *Conclusion*), in order to label explicitly all the meaningful aspects of the scientific discourse.

Such rhetoric characterization of markup structures can be specified through *DoCO*, the *Document Components Ontology*[20], and *DEO*, the *Discourse Elements Ontology*[21]. The following excerpt expresses that the elements *div*, *h3* and *p*, introduced in the previous excerpt, represent, respectively, a *section* of the paper (*#10*), a *section title* (*#11*), and a *paragraph* (*#12*) introducing some *background* assets (*#13*):

```
:rhetoric { :div la:expresses doco:Section . #10
:h-sec-3-1 la:expresses doco:SectionTitle . #11
:p1-sec-3-1 la:expresses doco:Paragraph , #12
  deo:Background } #13
```

Besides its structural and rhetorical characterisation, a document takes also part to a *citation network* with its cited documents, in particular taking into account the *reasons* for particular citations – e.g. to express qualification of or disagreement with the ideas presented in the cited paper – which may significantly effect the evaluation of a citation network itself.

For instance, analysing the content of the paper, for instance the aforementioned 1st paragraph of the 3rd section (1st subsection) of the paper (i.e. *:p1-sec-3-1*), we encounter several citations to other works that are introduced for a particular reason, e.g. to express qualification of or disagreement with the ideas presented in the cited papers. Using CiTO, the *Citation Typing Ontology*[22] [23], we provide descriptions of the nature of the citations, as shown in the following example, where paper "[5]" is cited as a source of background information (*#14*), and paper "[7]" is also cited as evidence supporting a statement (*#15*):

[20] DoCO: `http://purl.org/spar/doco`
[21] DEO: `http://purl.org/spar/deo`
[22] CiTO: `http://purl.org/spar/cito`

```
:citation { :delosdlms
  # citation to [5] in Sec 3.1, Par 1
  cito:obtainsBackgroundFrom #14
    <http://doi.ieeecomputersociety.org/10.1109/ICME
        .2005.1521528> ;
  # citation to [7] in Sec 3.1, Par 1
  cito:citesAsEvidence #15
    <http://dx.doi.org/10.1109/ICCV.1998.710779> }
```

In addition, strictly correlated with the citational aspects of a document, we can detail the organization of the claims and the arguments of the paper (providing evidences to a claim). The argumentative organisation of discourse is described using *AMO*, the *Argument Model Ontology*[23], that implements Toulmin's model of argumentation [11][24] in OWL, as shown in Fig. 1 and introduced in the following excerpt:

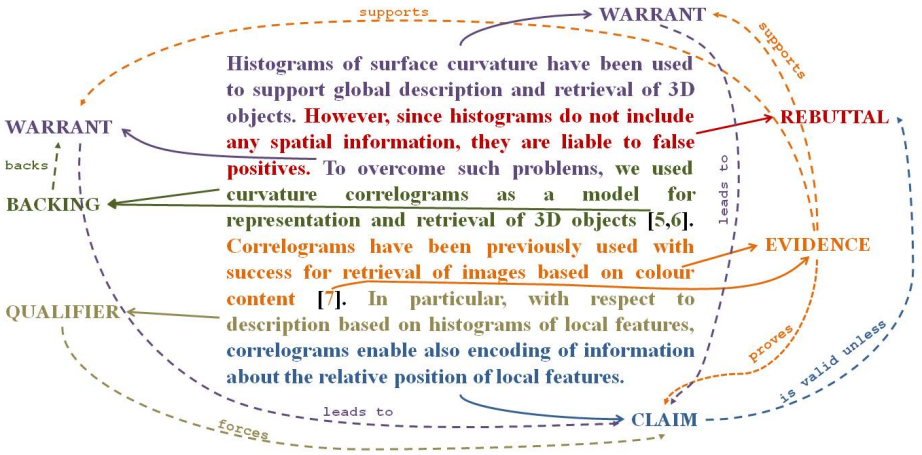


Fig. 1. Argument organisation of the 1st paragraph of Sec. 3.1 in the DelosDMS paper

```
:argumentation { :argument a amo:Argument ;
  amo:hasClaim :r-cl-p1 ; # correlograms...about
  amo:hasQualifier :r-qual-p1 ; # In...features
  amo:hasRebuttal :r-reb-p1 ; # However...false positive
  amo:hasEvidence :r-ev-1-p1 ... # Correlograms...content
```

[23] AMO: `http://www.essepuntato.it/2011/02/argumentmodel`

[24] Toulmin proposed that arguments are composed of statements having specific argumentative roles: the *claim* (a fact that must be asserted), the *evidence* (a foundation for the claim), the *warrant* (a statement bridging from the evidence to the claim), the *backing* (credentials that certifies the warrant), the *qualifier* (words or phrases expressing the degree of certainty of the claim) and the *rebuttal* (restrictions that may be applied to the claim).

```
  amo:hasWarrant :r-war-2-p1 ... # To...problems
  amo:hasBacking :r-back-1-p1 ... # we used...3D objects
:r-qual-p1 amo:forces :r-cl-p1 .
:r-cl-p1 amo:isValidUnless :r-reb-p1
:r-ev-1-p1 amo:proves :r-cl-p1 ; amo:supports :r-war-1-p1 ...
:r-war-2-p1 amo:leadsTo :r-cl-p1 .
:r-back-1-p1 amo:backs :r-war-2-p1 . ...
```

Finally, the *textual semantics*, i.e. the very message contained in a piece of text, is the final step in the definition of the semantics of a piece of text. For instance, the formal description of a claim needs to be expressed in such a way as to represent as faithfully as possible the meaning of the claim itself. Since each document expresses content in domains that are specific of the topic of the paper, we do not seek to provide an encompassing ontology to express claims. In some cases, the claim of an argument can be encoded through using a simple model, e.g. DBPedia [29], as shown in the following excerpt. In other cases, an appropriate specific ontology for the domain might be chosen freely.

```
:semantics {dbpedia:Correlogram a dbpedia:Mathematical_model}
```

4 Conclusions

The evolution of modern digital collections implies that the metadata we associate to their content are enhanced and enriched with more and more information. Simple metadata model may increase the likelihood of their adoption, but eventually result in simple annotations and possibly in errors and misrepresentations of the associated documents. Modern publishers are now approaching digital publishing from a semantic perspective (aka *semantic publishing* [4]).

In this paper we verified our *semantic lenses* [5] to semantically enhance a published scholarly article with direct, explicit, and hopefully correct annotations about the context, structure and argumentation of the paper as well as its actual content. Since one of the criteria for evaluating digital libraries as complex systems is the performance, which "depends strongly on the formats, structure and representations of the content" [30], we strongly believe that the use of semantic lenses as ontological keys could markedly improve usefulness of a library of scholarly articles. We are now working on *Through A Lens* (*TAL*), a prototypical application[25] we developed as proof of concept of the use of semantic lenses in a real-case scenario, that enables the navigation and understanding of a scholarly document through these semantic lenses. We are now analysing the outcomes of a user testing session we undertook to demonstrate the efficacy of TAL when addressing tasks requiring deeper understanding and fact-finding on a document. Finally, along the lines of our previous work [28], we plan to develop automatic and semi-automatic approaches – based on ML and NLP techniques – for the enrichment of documents according to semantic lenses.

[25] Available at `http://www.essepuntato.it/2013/tal/LensedMika.html`

References

1. Peroni, S., Tomasi, F., Vitali, F.: Reflecting on the Europeana Data Model. In: Proceedings of the 8th Italian Research Conference on Digital Libraries. Revised Selected Papers, pp. 228–240 (2012)
2. Whalley, W.B., MacNeil, J., Munroe, G., Landy, S., Power, S.: Developing a flexible structure for a pure e-journal. In: Rowland, F., Meadows, J. (eds.) Proceedings of the 1st ICCC/IFIP Conference on Electronic Publishing: New Models and Opportunities (1997), http://elpub.scix.net/data/works/att/97112.content.pdf
3. Dublin Core Metadata Initiative. DCMI Metadata Terms. DCMI Recommendation (June 14, 2012), http://dublincore.org/documents/dcmi-terms/
4. Shotton, D., Portwin, K., Klyne, G., Miles, A.: Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. PLoS Computational Biology 5(4), e1000361 (2009), doi:10.1371/journal.pcbi.1000361
5. Peroni, S., Shotton, D., Vitali, F.: Faceted documents: describing document characteristics using semantic lenses. In: Proceedings of the ACM Symposium on Document Engineering, pp. 191–194. ACM, New York (2012), doi:10.1145/2361354.2361396
6. Agosti, M., Berretti, S., Brettlecker, G., Del Bimbo, A., Ferro, N., Fuhr, N., Keim, D.A., Klas, C., Lidy, T., Milano, D., Norrie, M.C., Ranaldi, P., Rauber, A., Schek, H., Schreck, T., Schuldt, H., Signer, B., Springmann, M.: DelosDLMS - The Integrated DELOS Digital Library Management System. In: Proceedings of the 1st International DELOS Conference, pp. 36–45 (2007), doi:10.1007/978-3-540-77088-6_4
7. Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., Attwood, T.K.: Ceci nest pas un hamburger: modelling and representing the scholarly article. Learned Publishing 24, 207–220 (2011), doi:10.1087/20110309
8. Darcus, B., Giasson, F.: Bibliographic Ontology Specification. Specification Document (November 4, 2009), http://bibliontology.com/specification
9. IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records (FRBR). Final Report (1998), http://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm
10. Groza, T., Möller, K., Handschuh, S., Trif, D., Decker, S.: SALT: Weaving the claim web. In: Aberer, K., et al. (eds.) ISWC/ASWC 2007. LNCS, vol. 4825, pp. 197–210. Springer, Heidelberg (2007)
11. Toulmin, S.: The uses of argument. Cambridge University Press, Cambridge (1959)
12. Lauritsen, M., Gordon, T.F.: Toward a general theory of document modeling. In: Proceedings of the 12th International Conference on Artificial Intelligence and Law, pp. 202–211 (2009), doi:10.1145/1568234.1568257
13. Schneider, J., Groza, T., Passant, A.: A review of argumentation for the Social Semantic Web. Semantic Web -Interoperability, Usability, Applicability (2012), Pre-press available at: http://iospress.metapress.com/content/016x47v66347462v/fulltext.pdf (last visited July 25, 2012) (in press)
14. Koutsomitropoulos, D.A., Solomou, G.D., Alexopoulos, A.D., Papatheodorou, T.S.: Semantic Web enabled digital repositories. International Journal on Digital Libraries 10, 179–199 (2009), doi:10.1007/s00799-010-0059-z
15. Ranganathan, S.R.: Theory of Library Catalogue. Madras Library Association, Chennai (1938)

16. Quintarelli, E.: Folksonomies: power to the people. Presented at the ISKO Italy-UniMIB meeting (2005), http://www.iskoi.org/doc/folksonomies.htm
17. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. W3C Working Draft (03 May 2012). World Wide Web Consortium (2012), http://www.w3.org/TR/prov-o
18. Lee, C.A.: A framework for contextual information in digital collections. Journal of Documentation 67(1), 95–143 (2011), doi:10.1108/00220411111105470
19. Mazzini, S., Ricci, F.: EAC-CPF Ontology and Linked Archival Data. In: Semantic Digital Archives. In: Proceedings of the 1st International Workshop on Semantic Digital Archives (2011), http://ceur-ws.org/Vol-801/
20. International Council on Archives: ISAAR(CPF): International Standard Archival Authority Record for Corporate Bodies, 2nd edn. (2003), http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf
21. Pitti, D.: Creator Description: Encoded Archival Context. In: Taylor, A.G., Tillett, B.B., Baca, M., Guerrini, M. (eds.) Authority Control in Organizing and Accessing Information: Definition and International Experience, pp. 201–226. Haworth Information Press, Binghamton (2004)
22. Peroni, S., Shotton, D., Vitali, F.: Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents. In: Proceedings of the 8th International Conference on Semantic Systems, pp. 9–16. ACM, New York (2012), doi:10.1145/2362499.2362502
23. Peroni, S., Shotton, D.: FaBiO and CiTO: ontologies for describing bibliographic resources and citations. Journal of Web Semantics: Science, Services and Agents on the World Wide Web (2012), doi:10.1016/j.websem.2012.08.001
24. Text Encoding Initiative Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2005), http://www.tei-c.org/Guidelines/P5
25. Ore, C.E., Eide, O.: TEI and cultural heritage ontologies: Exchange of information? Literary and Linguistic Computing 24(2), 161–172 (2009), doi:10.1093/ll-c/fqp010
26. Di Iorio, A., Peroni, S., Vitali, F.: A Semantic Web Approach To Everyday Overlapping Markup. Journal of the American Society for Information Science and Technology 62(9), 1696–1716 (2011), doi:10.1002/asi.21591
27. Peroni, S., Gangemi, A., Vitali, F.: Dealing with Markup Semantics. In: Proceedings the 7th International Conference on Semantic Systems, pp. 111–118. ACM, New York (2011), doi:10.1145/2063518.2063533
28. Di Iorio, A., Peroni, S., Poggi, F., Vitali, F.: A first approach to the automatic recognition of structural patterns in XML documents. In: Proceedings of the 2012 ACM Symposium on Document Engineering, pp. 85–94. ACM, New York (2012), doi:10.1145/2361354.2361374
29. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web 7(3), 154–165 (2009), doi:10.1016/j.websem.2009.07.002
30. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C., Kovacs, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Solvberg, I.: Evaluation of digital libraries. International Journal on Digital Libraries 8(1), 21–28 (2007), doi:10.1007/s00799-007-0011-z

Digital Archives: Extending the 5S Model through NESTOR

Nicola Ferro and Gianmaria Silvello

Department of Information Engineering, University of Padua, Italy
{ferro,silvello}@dei.unipd.it

Abstract. Archives are an extremely valuable part of our cultural heritage. Although their importance, the models and technologies that have been developed over the past two decades in the *Digital Library (DL)* field have not been specifically tailored on archives and this is especially true when it comes to formal and foundational frameworks, as the *Streams, Structures, Spaces, Scenarios, Societies (5S)* model is. Therefore, we propose an innovative formal model, called *NEsted SeTs for Object hieRarchies (NESTOR)*, for archives, using it to extend the 5S model in order to take into account the specific features of the archives and to tailor the notion of digital library accordingly.

1 Motivation

Over the past two decades, *Digital Libraries (DLs)* have been steadily evolving and have been shaping the way in which people and institutions access to and interact with our cultural heritage, study, and learn. Nowadays, their reach goes far beyond what has been the realm of traditional libraries and encompasses also other kinds of cultural heritage institutions, such as archives and museums. In particular, this work focuses on archives; an archive is not simply constituted by a series of objects that have been accumulated and filed with the passing of time – as it usually happens with libraries that collect, for example, individual published books, journals, and serials. Instead, it represents the trace of the activities of a physical or juridical person in the course of their business which is preserved because of their continued value.

DLs benefit from the existence of sophisticated formal models, such as the *Streams, Structures, Spaces, Scenarios, Societies (5S)* model [4], which allow us to formally describe them and to prove their properties and features. Notwithstanding the importance of the archives, so far, there has been no attempt to develop a dedicated formal model, built around their peculiar constituents, such as the notion of *archival bond*. We can neither exploit the 5S model as it is for archives because, as we will discuss later on, it needs some kind of extension and tailoring.

We think that the archival domain deserves a formal theory as well and that this theory has to be reconciled with the more general theories for digital libraries in order to disclose to archives the full breadth of methodologies and technologies

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 130–135, 2014.

which have been developed over the last two decades in the DL field. To this purpose we proposed a formal model for archives, built around the notion of *archival bond* and *hierarchy*: the *NEsted SeTs for Object hieRarchies (NESTOR)* model [1]. Furthermore, we exploit NESTOR to formally extend the 5S model in order to be capable of defining a *digital archive* as a specific case of digital library able to take into consideration the peculiar features of the archives.

The paper is organized as follows: in Section 2 we provide some background on archives and the 5S formal model. In Section 3 we present the basics of the NESTOR model and in Section 4 we introduce our extension to the 5S model via NESTOR. Finally, in Section 5 we draw some final remarks.

2 Related Work

2.1 Digital Archives

In an archive the context and the relationships between the documents are preserved thanks to the hierarchical organization of the documents inside the archive. Indeed, an archive is divided by fonds and then by sub-fonds and then by series and so on; at every level we can find documents belonging to a particular division of the archive or documents describing the nature of the considered level of the archive. The union of all these documents, the relationships and the context information permits the full informational power of the archival documents to be maintained. The archival documents are analyzed, organized, and recorded by means of the *archival descriptions* that have to reflect the peculiarities of the archive. In the digital environment archival descriptions are encoded by the use of metadata; these need to be able to express and maintain the structure of the descriptions and their relationships [3].

The standard format of metadata for representing the hierarchical structure of the archive is the *Encoded Archival Description (EAD)*[1], which reflects the archival structure and holds relations between entities in an archive. On the other hand, an archive is described by means of a unique EAD file and this may be problematic when we need to access and exchange archival metadata with a variable granularity [2].

2.2 The 5S Model

The *Streams, Structures, Spaces, Scenarios, Societies (5S)* [4] is a formal model and draws upon the broad DL literature in order to have a comprehensive base of support. It has been developed largely bottom up, starting with key definitions and with elucidation of the DL concepts from a minimalist approach. It is built around five main concepts:

- *streams* are sequences of elements of an arbitrary type, e.g. bits, character, images, and so on;

[1] `http://www.loc.gov/ead/`

- *structures* specify the way in which parts of a whole are arranged or organized, e.g. hypertexts, taxonomies, and so on;
- *spaces* are sets of objects together with operations on those objects that obey certain constraints, e.g. vector spaces, probabilistic spaces, and so on;
- *scenarios* are sequences of related transition events, for instance, a story that describes possible ways to use a system to accomplish some functions that a user desires;
- *societies* are sets of entities and relationships between them, e.g. humans, hardware and software components, and so on.

Starting from these five main concepts, it provides a definition for a minimal DL which is constituted by: (i) a repository of digital objects; (ii) a set of metadata catalogs containing metadata specifications for those digital objects; (iii) a set of services containing at least services for indexing, searching, and browsing; and, (iv) a society.

While these broad concepts can be in common also with archives, when you look at the specific way in which they are formally defined, you realize that the definitions cannot be straightforwardly applied to the archives case without at least some extension. We will discuss this in further details, presenting an extension of 5S via NESTOR in Section 4.

3 The Basics of the NESTOR Formal Model

We define both *Nested Sets Model (NS-M)* and *Inverse Nested Sets Model (INS-M)* in terms of the set theory as a collection of subsets where specific conditions must hold.

Definition 1. *Let A be a set and let $\mathcal{C}$ be a collection of subsets of A. Then $\mathcal{C}$ is a* ***Nested Sets Collection*** *(NS-C) if:*

$$A \in \mathcal{C}, \tag{3.1}$$

$$\forall H, K \in \mathcal{C} \mid H \cap K \neq \emptyset \Rightarrow H \subseteq K \vee K \subseteq H. \tag{3.2}$$

Therefore, we define a NS-C as a collection of subsets where two conditions must hold. The first condition (3.1) states that set A which contains all the subsets of the collection must belong to the NS-C itself. The second condition states the intersection of every couple of sets in the NS-C is not the empty-set only if one set is a proper subset of the other one.

Now we can introduce the Inverse Nested Sets Collection (INS-C) which defines the INS-M:

Definition 2. *Let A be a set and let $\mathcal{C}$ be a collection. Then, $\mathcal{C}$ is an* ***Inverse Nested Sets Collection*** *(INS-C) if:*

$$\exists! B \in \mathcal{C} \mid \forall K \in \mathcal{C}, B \subseteq K, \tag{3.3}$$

$$\forall H, K, L \in \mathcal{C} \mid H \subseteq K, L \neq K \Rightarrow (L \cap K = H \cap L) \vee (H \subseteq L) \vee (L \subseteq H). \tag{3.4}$$

We define an INS-C as a collection of subsets where two conditions must hold. The first condition (3.3) states that $\mathcal{C}$ must contain the *bottom set* B, which is the common subset of all the sets in $\mathcal{C}$. The second condition (3.4) states that if we consider three sets K, H, and L such that H is a subset of K and K is not equal to L, then the intersection between L and K is not the same as the intersection between H and L or H is not a subset of L and vice versa.

4 Extending the 5S Model via NESTOR

The notion of *descriptive metadata specification*[2] (definition 14 [4, p. 293]) is suitable either to represent, for each archival division, a descriptive metadata – e.g. a metadata describing a serie, a sub-fonds, or an archival unit – or to represent the archive as a whole, as it happens in the case of EAD.

When it comes to the definition of *metadata catalog* (definition 18 [4, p. 295]), there is no means to impose a structure over the descriptive metadata in the catalog. Therefore, if you use separate *descriptive metadata specifications* for each archival division, as in the former case, this would prevent the possibility of expressing the relationships between these archival divisions, i.e. you would loose the possibility of retaining the archival bond.

Moreover, in a *metadata catalog*, there is no means to associate (sub-)parts of the *descriptive metadata specifications* to the *digital objects* (definition 16 [4, p. 294]) that they describe, but you can only associate a whole descriptive metadata to a whole digital object.

Therefore, if you represent an archive as a whole with a single *descriptive metadata specification*, as in the latter case, it would not be possible to associate (sub-)parts of that descriptive metadata to the different digital objects corresponding to the various archival divisions.

Our extension to the 5S model is thus organized as follows:

- using the notion of *structure* (definition 2 [4, p. 288]), we introduce the notion of **NESTOR structure**, as a structure that complies with the constraints of NS-M or INS-M;
- using the notion of *metadata catalog*, we introduce the notion of **NESTOR metadata catalog**, as a metadata catalog that exploits a NESTOR structure to retain the archival bonds;
- using the notion of *digital library* (definition 24 [4, p. 299]), we introduce the notion of **digital archive**, as a digital library where at least one of the *metadata catalogs* is a NESTOR metadata catalog.

Definition 3. *Let $\mathcal{C}$ be a Nested Set Collection (NS-C) on a set A. A* ***NS-M structure****(A) is a structure $(NS\text{-}G, L, \mathcal{F})$, where L is a set of label values, $\mathcal{F}$ is a labeling function, and $NS\text{-}G = (V, E)$ is a directed graph where $\forall v_j \in V, \exists!\ J \in \mathcal{C} \wedge \forall e_{j,k} \in E, \exists!\ J, K \in \mathcal{C} \mid K \subseteq J$.*

[2] In this section, we use italics for highlighting definitions taken from the 5S model.

Definition 4. *Let $\mathcal{C}$ be an Inverse Nested Set Collection (INS-C) on a set A. A* ***INS-M structure****(A) is a structure $(INS\text{-}G, L, \mathcal{F})$, where L is a set of label values, $\mathcal{F}$ is a labeling function, and $INS\text{-}G = (V, E)$ is a directed graph where $\forall v_j \in V, \exists! J \in \mathcal{C} \wedge \forall e_{j,k} \in E, \exists! J, K \in \mathcal{C} \mid J \subseteq K$.*

Definition 3 applies definition 1, ensuring that the resulting structure complies with the NS-M. Note that the set of label values L and the labeling function $\mathcal{F}$ are not strictly needed for the NS-M, but they can be useful in the context of the 5S and this feature, in turn, may extend the NS-M with semantic possibilities. Similarly, definition 4 applies definition 2.

Definition 5. *Given a set A, a* ***NESTOR structure****(A) is either a NS-M structure(A) or a INS-M structure(A).*

The definition of *metadata catalog* in the 5S model can be expressed as follows. Let H be a set of handles to *digital objects* and M a set of *descriptive metadata specifications*, then a *metadata catalog* is a function $DM : H \times 2^M$.

Definition 6. *Let H be a set of handles to digital objects and M a set of descriptive metadata specifications, a metadata catalog DM is a* ***NESTOR metadata catalog*** *if:*

$$\forall h_i \in H \mid \exists M_i \in 2^M \ \wedge \ DM(h_i) = M_i \Rightarrow |M_i| = 1 \tag{4.1}$$

$$\exists \text{ NESTOR structure}(M) \tag{4.2}$$

Condition 4.1 imposes that, if exists, there is only one *descriptive metadata specification* for a given *digital object* because, in the archival practice, every single metadata describes a unique archival division, being it a level in the archive or a digital object [5]. Condition 4.2 ensures that the relationships among the different archival divisions are compliant with the *descriptive metadata specifications* in M.

Definition 7. *A* ***digital archive*** *$(\mathcal{R}, DM, Serv, Soc)$ is a digital library where*

- *$\mathcal{R}$ is a repository;*
- *at least one of the metadata catalogs in the set of metadata catalogs DM is a NESTOR metadata catalog;*
- *Serv is a set of services containing at least services for indexing, searching, and browsing;*
- *Soc is a society.*

Definition 7 extends the definition of *digital library* in the 5S model requiring that at least one of the *metadata catalog* is a NESTOR one, i.e. there exists at least on *metadata catalog* capable of retaining the archival bonds.

5 Final Remarks

The definition of digital archive we gave in this paper has a couple of consequences. Firstly, more NESTOR metadata catalogs can be present in the same digital archive, thus giving the possibility of expressing different archival descriptions over the same set of *digital objects*. This extends the current practice in which a system for managing an archive is usually capable of managing only one description of the archive, thus giving only one point-of-view on the held material. Secondly, you can mix NESTOR and not-NESTOR metadata catalogs which allows for seamlessly integration of different visions of the managed *digital objects* within the same digital archive. This opens up the possibility of exploiting the whole breadth of methodologies and tools available in the DL field with the archives.

Future work will concern the formal definition of creation, deletion, update, and search operations on digital archives via NESTOR. This, in turn, will open up the possibility to further extend the 5S model. Indeed, according to it, a minimal digital library has to offer, at least, indexing, searching, and browsing services [4, p. 299].

Acknowledgments. CULTURA[3] (Grant agreement no. 269973) and the PROMISE network of excellence[4] (Contract n. 258191) projects, as part of the 7th Framework Program of the European Commission, have partially supported the reported work.

References

1. Agosti, A., Ferro, N., Silvello, G.: The NESTOR Framework: Manage, Access and Exchange Hierarchical Data Structures. In: Proceedings of the 18th Italian Symposium on Advanced Database Systems, pp. 242–253. Società Editrice Esculapio, Bologna (2010)
2. Ferro, N., Silvello, G.: A Methodology for Sharing Archival Descriptive Metadata in a Distributed Environment. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 268–279. Springer, Heidelberg (2008)
3. Gilliland-Swetland, A.J.: Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment. Council on Library and Information Resources, Washington, DC, USA (2000)
4. Gonçalves, M.A., Fox, E.A., Watson, L.T., Kipp, N.A.: Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems (TOIS) 22(2), 270–312 (2004)
5. International Council on Archives. ISAD(G): General International Standard Archival Description, 2nd edn. International Council on Archives, Ottawa (March 1999)

[3] http://www.cultura-strep.eu/
[4] http://www.promise-noe.eu/

Evaluation of Digital Humanities: An Interdisciplinary Approach

Anna Maria Tammaro

University of Parma, Department of Information Engineering

Abstract. This research, now in its first phase of development, focuses upon evaluation of Digital Humanities, here indicated as the disciplines, included within the Italian Disciplinary Areas 10 and 11, innovating their research outputs through the application of technological methods. These research outputs are relevant both quantitatively and qualitatively, but do not seem to be considered of value by the current procedures of quality evaluation. The research methodology is including a comparison of the evaluation policies and quality assurance procedures in Europe regarding the different typologies of digital publications. The final product will be a KOS (Knowledge Organization System) based on the Web standards, such as RDF and Linked Open data, to represent and organize the digital products and publications as well as the related agents (persons, institutions, etc.). The KOS will include the results derived by user studies including: 1) a toolkit that will provide rich and meaningful information about the research activity and publications in Digital Humanities. The toolkit will consist of decision tools, able to analyze the content of the proposed knowledge organization system; 2) a platform for the diffusion of Project results, including digital publications, OER for training and other communication tools.

Keywords: Digital Humanities, Bibliometrics, Peer review, Digital publishing.

1 Introduction

The definition of criteria which can identify the quality of scientific production is an issue of important national and international interest. Such evaluation is necessary both for ranking by quality the competitiveness of universities and of the nations, as well as - on the individual level - for the recruitment and tenure of a single scholar. In recent years the focus on research assessment has grown exponentially in Italy, as a proportion of resources devoted to research is now allocated on the basis of the results of the national research assessment exercises. In July 2009 the Ministry of Education and Research has for the first time allocated funds to the universities referring in part to the VTR (Triennial Evaluation of Research), the first national research assessment exercise which ended in 2007 (Baccini, 2010). Italy is currently carrying out its second national research assessment exercise, called VQR (Evaluation of the Quality of Research, 2004-2010).

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 136–146, 2014.

Traditionally the evaluation of scientific publications in the humanities disciplines is based upon "peer review". This is coherent with the prevalent definition of "scientific base" which, as was noted by the CUN (National Unversity Council 2010), essentially rests upon the affiliation and consensus of a particular scientific community. One problem in evaluation based upon peer review is that this can be based upon "subjective" judgments, not avoiding conservative and corporate trends. Peer review also hinges upon the publishing system of scientific publications, which is completely controlled by commercial publishers, who have assumed the role of guarantors for the quality of academic publications. The accreditation of quality is done by publishers through the publishing process: the commercial publishers have the monopoly of the peer review, also if the reviewers are scholars in public institutions. The current evaluation procedure established by ANVUR (National Agency Evaluation of the University Research), continues to base itself upon peer review (effected by Experts Group Evaluation - GEV) but, in order to limit the risks of quality evaluation based upon peer review, it has added bibliometric indicators, such as h-Index and Impact Factor, to have "objective" evaluation tools for publications produced in the humanities. The ANVUR procedure however continues to consider the publisher as the only referent for the quality of the editorial flow of the scientific publications. The peer review and the impact indicators are indeed not so different as many declare: both are based on the same closed scholarly community and continue to focus on traditional publications channels.

Scholarly communication is however using alternative publishing models using the Internet as main communication channel. These online publications include Open Access publications, blogs, wiki, RSS, Twitter and other online publications. The characteristic of these parallel publications models in the humanities is that they are marked, quickly updated, multimedia and hypertextual but they do not follow the traditional editorial flow (Fig. 1).

The CUN (National University Council, 2010) established that "sono ammessi i prodotti di ricerca non aventi natura di pubblicazioni purchè corredati da documentazione atta a consentire le valutazione" ["research works which do not have the same nature as printed publications are permitted provided that they are furnished

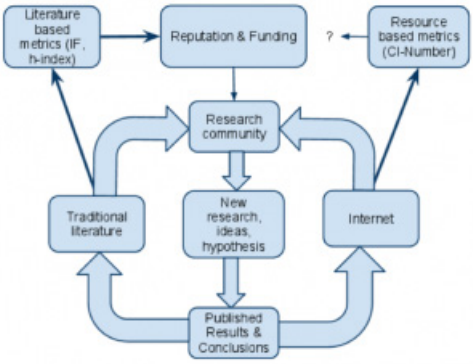


Fig. 1. Models of publication in the humanities

with documentation which permits their evaluation"]. Even the ANVUR procedures continue to ignore online publications. Regarding digital publications, ANVUR is considering only the version in PDF format of printed publications / hard copies.

Thus it would seem that the current evaluation process ignores Internet products and digital publications for evaluation both of institutions and individual. This fact can be interpreted as an overall judgment of publications without quality and thus not even worthy of being evaluated.

Digital Humanities, here indicated as the disciplines included within the Disciplinary Areas 10 and 11, are now innovating their research behaviour and outputs through the application of technological methods. The advent of digital technologies has produced a completely new scenario, starting from pioneers such as Antonio Zampolli and padre Busa, where the humanities should start to rethink one's own horizon of disciplinary products of research and the tools to use them.

Philosophers of communication and cognitive psychologists have been investigating for years now on the new semiotic system introduced by the emergence of the new digital publications (which are marked, multimedia, hypertextual), even the historians of contemporary literature and especially philologists have been called to the new challenges and to consider the fact that the digital revolution impacts on the critical analysis of texts. Digital philology, for example, gives the possibility to offer not only the text prepared by the philologist, but also variants rejected, together with the tools offered by the "computational philology" for studying the "corpora". The example shows that the research done by the Digital Humanities groups has an important characteristic: it is interdisciplinary. The interdisciplinarity however is an added particular obstacle for traditional evaluation in Italy, when it involves scholars from the humanistic Areas together with scholars from the Information Technology Area, with different quality assurance procedures: what is the Area who can better evaluate this research?

2 Aims and Objectives

The problem from which the research project gets underway is: there are new genres of publications in the humanities and new models of communication using the Internet. Can we evaluate digital products and publications with current quality indicators and the same procedures of printed publications? Do we perhaps need to have an alternative system for the evaluation of digital products and publications?

While internationally digital products and publications must be treated more or less by the same standards as printed publications, the evidence in Italy is that, by not assuming that these are a form of publication, they are not even considered (Tammaro 2002; Huang Chang 2008; OECD 2008; White 2009; Fister 2010; Hellqvist 2010). Many of these outputs have no visibility, as they are not registered in current registries, catalogues and bibliographies, starting from the Italian Research Registry (Anagrafe della Ricerca).

The aim of this research is to give transparency to criteria which are currently used in Italy and abroad for the evaluation of digital publications. The project aims to

improving international visibility of Italian scientific production in the Digital Humanities and its originality consists in trying to understand the current model of scientific communication in Italy that is taking shape for products and publications in the Digital Humanities.

The objectives are:

1. collaborating with international institutions and partners to realise an ontology, and
2. ensuring interoperability with tools developed abroad.

To achieve this goal we intend to use the methodology of the Open Linked Data and Semantic Web, proposing a tool that aggregates and classifies digital publications in the field using citations and quantitative indicators and in the same time enabling an open peer review. The first objective is to propose a tool which aggregates, classifies using Open Data such as citations and other quantitative and qualitative indicators and makes an open peer review possible.

Other objectives of the project are further enhancement of the international network of research achieved with the participation of qualified experts from Italy and different countries, and promoting advanced training for staff, academics and young researchers.

2.1 Typologies of Digital Publications

This research intends to point out the different typologies of digital publications existing in the humanities reaching their overall classification. In this research by digital products and publications we mean not only publications which have an equivalent print version, but even those which are only digital and which, in the restricted meaning adopted by CUN, might not be considered publications.

For example: is an e-learning course a digital publication? Is the documentation of a software, a database or a Web site which is continuously updated, a digital publication? Is a critical text analysis combined with different variants of text a publication? Many of the digital products are not only used for research but also for teaching, often encouraged by universities which have often made a platform for e-learning available.

The publications in Digital Humanities are the result of research and teaching and include different types, some corresponding to digitised publications with a print version, but the majority are new types, such as marked texts, data sets and dynamic databases, wikis, RSS, Web sites. In addition we must point out that the traditional system of quality evaluation does not adapt to digital publications which adopt completely different multimedial systems and editorial processes. The digital publications are hypertextual, dynamic, easily accessible, can be open and are re-usable. Digital publications do not necessarily follow a process entirely managed by the editor: many are made available in Open Access modes, in institutional repositories of universities, on the university department Web sites, in Open Access journals, in University series which are only available online. Digital publications are also stimulating an ongoing process of research, analysis and collaborative work over a final and fixed product. This creates a fundamental challenge for the review of new model scholarship.

The gap which is evident is that the traditional system of quality evaluation does not adapt to these new genres of digital products and publications, which adopt completely different multimedial systems and publishing processes. How does the evaluation of research in the digital environment change? Do the disciplinary fields which control the quality of research follow both formal and informal criteria which can also be used for digital publications? Are there cultural barriers which hinder an efficient evaluation of digital publications?

Digital products and publications are often open in the Web, but paradoxically they are not often registered in scientific bibliographies and catalogues and other finding aids. Digital publications, due to their characteristics of hypertextual and open format, are also interdisciplinary and belong to different Disciplinary Areas such as those of Area 9 related to Information Engineering, which in this Project has been included as the Area of the comparative analysis for the evaluation criteria and processes.

The purpose of this research is to understand the values and purposes of the criteria currently used in Italy and to compare them with the European context for evaluating digital publications so as to better identify the quality of these publications in the humanities disciplines. Because of their characteristics, digital products and publications offer the possibility for applying a new methodology of collective and open evaluation, which combines both the peer review and the bibliometric indexes.

3 Literature Review

These new publications and digital assets have acquired a great importance for the new opportunities that they give to advancing research and teaching in the humanities and their characteristics were discussed by the scientific community nationally and internationally (Mordenti 1987; Buzzetti 1999; Tammaro 2001; Mordenti 2001, Robinson 2005; Buzzetti, 2006; Mordenti 2006, Perrault 2006; Dyes 2007; Roncaglia, 2008; Orlandi 2010; Deegan & Tanner 2005).

In the professional literature, as well as in the evaluation procedures, it is traditionally made a very clear distinction between qualitative (peer review) and quantitative assessment (bibliometrics). Peer review and bibliometric measures are two methodologies widely used in the field of hard sciences.

With regard to the qualitative assessment, it is based on the methodology of the peer review, which in the case of publications and research projects is usually conducted ex-ante, while in the national research assessment exercises such as ANVUR is normally carried out ex-post. The quantitative methodology used for research assessment relies, on the contrary, on impact indicators. The best known are the bibliometric indicators, and among these, the most used are the citational indicators (Impact Factor, H-Index and its variants, Eigenfactor etc.). Bibliometric measures are methodologies widely used in the field of hard sciences, also if sometimes abused, and criticized for this. They are, however, less practiced in the field of the Humanities and Social Sciences, where scholars put into question their validity and applicability. Moreover in the Humanities the Impact Factor is never calculated.

Due to the specific characteristics of digital publications, the evaluation procedure of digital publications have stimulated the development and experimentation of a transparent, open and collective quality evaluation, which combines qualitative and quantitative systems evidencing the scientific base of digital content (Tammaro 2001; Reale 2008; Guerrini 2009; O Rieger 2010). Some tools that are based on the analysis of citations have been applied to the humanities as for example: SCImago. (2007). SJR1 - SCImago Journal & Country Rank is being developed by a consortium of Spanish universities (Granada, and Madrid Alcal Charles III). It 's based on data from Elsevier's Scopus database combined with the algorithm of Google's page rank. It can get a list in order of importance (visibility) of the periodic and partner countries. Latindex2 focuses on the scientific journals of Spain, Portugal and Latin America. The basic idea is to have a unique system for evaluating scientific journals regardless of the medium with the addition of some specific indicators for electronic journals. The consortium that runs it has produced research for the e-book and electronic publications in Open Access. For the comparison of large amounts of data, as for the evaluation of departments and universities, the bibliometric systems have interesting results. In 2006 the network Publish or Perish (PoP) started to offer access to a free-ware program developed by Anne-Wil Harzing, Professor of International Management at the University of Melbourne, Australia (www.harzing.com). Based on Google Scholar, this program provides in a few seconds the main bibliometric variables in all major fields of research, from management to the social sciences, through the "hard" sciences. Be considered for the evaluation of digital publishing tools that are based on specific statistical algorithms research that consider the number of accesses, downloads and views. In 2004 Alex Verstak and Anurag Acharya, two engineers working on Google have launched Google Scholar, a search engine that directs, with the agreement of the publishers, the entire text of scientific articles of a large number of scientific publications covering all disciplines. Google scholar applies the same algorithm of ranking Web pages in the work of researchers and you can then make comparisons within the same field of research. The tools of Social bookmarks such as refworks.com, zotero.org, connotea.org, mendeley.com, 2collab.com, citeulike.org, mekentosj.com, are considered as collaborative assessment tools that help in evaluation.

The limitation of these tools is that we know of each Web page the number of visits but this should not be confused with the impact. To overcome this problem, others have used the tools that combine the system of citations with the peer reviews, publishing the results online and open to the rear. Mesur (Metrics for Scholarly Usage of Resources), funded by the Mellon Foundation in 2006 has attempted to bring together bibliographic information, citations and data used to create a working model of scientific communication. In the first phase an ontology was developed that in the second phase was applied to data extracted by the project partners in collaboration with some editors and the project COUNTER (Bollen, Van de Sompel, Rodriguez 2006).

Recently, social media, such as Twitter, Mendeley, Google Groups, FriendFeed and Linkedin, have been used for analysing and filtering digital publications and informing scholarship. Impactstory and ScienceCard are Web based applications to track the impact of a wide range of research publications and of individual scholars.

However, there are obstacles for the evaluation of digital publications, which have been highlighted for example by Vanhoutte (2006) and that are related to the difficulty of identification in the catalogs and bibliographic databases (Torres Salinas & Moed, 2009: White et al., 2009) and lack of the editorial process (pre-peer review). It therefore seems to us that we need tools and registries for the identification and classification of digital publications that are supporting the assessment procedures, both pre and post-publication. These tools need to include all the measurements of impact that we have now:

Peer review, as expert judgement;

Citations as impact factors;

Usage as downloads and views;

Alt-metrics considered as bookmarks, links, conversations in blogs, Twitter.

The new tools, including usage counts and alt-metrics, have a common characteristic: they extend the evaluation and availability of digital publications to an open research community and other stakeholders. Transparency of reviewing is the added value of these tools.

4 Methodology and First Findings

To achieve its objectives, the Project has created a team composed of research units with multidisciplinary skills from research institutes, universities, publishers companies, collaborating with experts from the community of the Associazione Italiana Informatica Umanistica e Cultura Digitale (AIUCD).

The Project team is considering in particular the digital publications in Open Access institutional repositories and digital publications available in Open Access mode with commercial publishers and digital libraries, publications and products related to research and teaching, data sets and all the other types considered important by the experts participating in the Project. An in-depth examination is planned in the first phase of the specific characters of three kinds of publications: electronic scholarly editions (ecdotique and digital philology), eBooks, learning materials made available as OER (Open Educational Resources). This work has been built on top of theoretical models developed until now in the field of Digital Libraries, in particular the DELOS/DL.org and the 5S ones, and, if necessary, the Zachmann Framework, related to general information system. The Project team is collecting data from all the existing dabases, bibliographies and catalogues existing in the Area.

A benchmarking has been done of some of the examples of tools cited before, such as those developed in the United States (Mesur, MERLOT), Spain (Scimago), France (OpenEdit) and UK (JISC 2009). The methodology proposed by the project, which has been used for example in the United States for projects such as MESUR and MERLOT and in France for the Open edition, seems to have a substantial impact on current procedures and criteria for the evaluation of digital publications and products in Italy, proposing a tool that aggregates bibliographies and catalogues and classifies digital publications and citations in the field using quantitative indicators, and enabling an open peer review.

A survey of experts opinions and experiences has started during the first Annual Conference of the AIUCD, with the creation of a Special Interest Group to continue the ongoing discussion. At the end of this phase, an international report on comparative analysis of the evaluation model of digital publications in the Digital Humanities in Europe has to be produced together with a report on the ontology of scientific communication and its application in a Knowledge organisation system (KOS). The products expected from this 3 years research are:

Analysis of the typologies of digital publications and of the criteria and procedures for their evaluation in Digital Humanities disciplines;

Construction of an ontology which evidences the junctions / hubs and relationships existing between digital publications in the Digital Humanities in Italy;

Construction of a Knowledge Organisation System, as a supporting tool for the choices of quality evaluation. The project will plan, develope and implement, within the Knowledge Organization System (KOS) of the project, a conceptual scheme (ontology) defined in a formal way and related to the different types of digital publication in the humanities, with a particular interest in the disciplinary Area 10 and 11, compared with those of Area 9.

The development phase also includes a possible formalization of DELOS/DL model, so to provide both a SPARQL endpoint, in order to allow for structured queries, and a publication of the RDF data following the Linked Data principles, in order to be included in the Linking Open Data Project Cloud (LOD), which is at the moment the most important concrete development of the Semantic Web.

The project also aims to produce and distribute the following products:

1. An international report on comparative analysis of the evaluation model of digital publications in the humanities in Europe;
2. An international report on the ontology of scientific communication and its application in KOS;
3. Development of open educational resources, to be used within the activities of diffusion and dissemination of research results. The proposted OER would illustrate the main typologies of electronic publications identified by the project and the characteristics of the main research metrics considered, aiming at contributing tools able to allow external user and groups the possibility of acquiring in a guided way the necessary skills for using both the theoretical and the practical tools produced.

The Project is still in the development phase and not started until now, due to an obstacle which has been met. The Evaluation Committee of the Area 11 who has evaluated it, has not financed it, with the following motivation:

"This will certainly be research with a high impact, mostly on the sector of research policy and research assessment; indirectly, it may also contribute to changing publication habits. How it can be viewed from a purely scholarly perspective is less clear because research assessment, bibliometry etc. is an emerging field - the impression is that the more scholarly (and therefore complex) it gets, the less will it be useful for actual quality control".

It seems to the author that the problem with the evaluation of publications is in the present closed research community of reviewers using both the peer review and the bibliometric methodologies. Instead the Digital Humanities publications need an extended community of evaluators using open tools.

5 Conclusion

This Project wants to indicate that some solutions can be found to the evaluation of digital products and publications by taking advantage of information and communication technology. In this area, in fact, the European Union countries are investing resources to improve the quality and visibility of research products. In particular, the Project considers the technologies of the Semantic Web a useful tool to facilitate the access of scholarly communities to quality research and provide expert reviewers with a tool to make more consistent and informed their decision making process, speed up the review, improve the procedures of peer review.

It is important to point out how the project aims not only to identify the technologies most suitable from a technical standpoint, but the focus is on developing a series of recommendations towards organizational, legal, and training needs, to support the diffusion of Digital Humanities research in a complex environment and contrasting the trend to be refractory to innovation such as in the humanities Area in Italy. The main recommendation is the need of an open research community for the evaluation of interdisciplinary products and publications.

This Project itself is a Digital Humanites project and covers both interdisciplinary dimensions. The scientific achievement is therefore twofold:

on the one hand, there is the need to identify, analyze and propose, some technical solutions;

on the other hand, it is an essential part to analyze and propose solutions to theoretical, organizational, educational and legislative issues to ensure a virtuous cycle of knowledge production.

The Project focuses on the design, implementation and testing of some applications, each of which is used for a specific methodological approach that will produce and disseminate knowledge on the use of technology in digital products and publications in Italy and Euroepan countries and, more generally, on the relationship between technology and organization of scholarly communication.

The paper considers the evaluation of digital products and publications in Digital Humanities in Italy a strategic objective to invest for innovation and competitiveness of research: there is the evidence of the need of extending the research community involved in the reviewing process to eliminate the present obstacle of “subjectivity”of the evaluation process. The "strategic value" for the country lies mainly in the need to find solutions to the complete lack or weakness of the evaluation for digital products and publications that are produced more and more numerous in the humanities, and renewing efforts to stimulate interdisciplinarity and avoiding the lack of visibility and recognition of excellence in research carried out in Italy for the Digital Humanities.

We must also say that investments in information technology and communication that all Italian universities have undertaken to provide an adequate infrastructure for research, do not seem to have affected the evaluation of quality of research products, stressing the need for project such as this, to explore new ways to find solutions to this interdisciplinary exploitation.

References

1. Baccini, A.: Valutare la ricerca scientifica: uso e abuso degli indicatori bibliometrici. Il mulino, Bologna (2010)
2. Barbieri, E.: La ricerca universitaria e la sua valutazione. Guaraldi, Rimini (2011)
3. Rodriguez, M.A., Bollen, J., Van de Sompel, H.: Mapping the bid behavior of conference referees. Journal of Informetrics 1(1), 68–82 (2006)
4. Buzzetti, D.: Archiviazione digitale dei dati e adeguatezza della rappresentazione del testo. Schede Umanistiche 13(2), 209–218 (1999)
5. Buzzetti, D.: Biblioteche digitali e oggetti digitali complessi: Esaustività e funzionalità nella conservazione. In: Archivi informatici per il patrimonio culturale, Atti del Convegno internazionale (Roma, Accademia Nazionale dei Lincei, Novembre 17-19), Roma, Bardi Editore (Contributi del Centro Linceo Interdisciplinare «Beniamino Segre», N. 114), pp. 41–75 (2003)
6. CRUI, Raccomandazioni. Open Access ed i prodotti della ricerca (2009), http://www.crui.it/Homepage.aspx?ref=1782
7. CUN (2010) Quattro anni di CUN per l'Università: 2007-2010, Roma, CUN (2010)
8. Deegan, M., Tanner, S.: Digital futures: strategies for the Information Age. Facet, London (2001)
9. Fister, B.: Getting Serious About Digital Humanities | Peer to Peer Review. LIbrary Journalcom (2010), http://www.libraryjournal.com/article/CA6729325.html (retrieved)
10. Guerrini, M., Ventura, R.: Problemi dell'editoria universitaria oggi: il ruolo delle university press e il movimento a favore dell'open access. In: Dalla pecia all'e-book: libri per l'università: stampa, editoria, circolazione e lettura: atti del convegno internazionale di studi: Bologna, Ottobre 21-25, a cura di Gian Paolo Brizzi, Maria Gioia Tavoni. CLUEB, Bologna (2009)
11. Hellqvist, B.: Referencing in the humanities and its implications for citation analysis. Journal of the American Society for Information Science and Technology 61(2), 310–318 (2010)
12. Chang, C.H., Ooi, G.L.: Role of Fieldwork in Humanities and Social Studies Education. In: Tan, O.S., McInerney, D.M., Liem, A.D., Tan, A.G. (eds.) Research in Multicultural Education and International Perspectives Series. What the West can learn from the East. Asian Perspectives on the Psychology of Learning and Motivation, vol. 7, pp. 295–312. Information Age Publishing, Charlotte (2008)
13. JISC, Meeting the research data challenge (2009), http://www.jisc.ac.uk/publications/briefingpapers/2009/bpresearchdatachallenge.aspx
14. Mordenti, R.: Su alcuni problemi di metodologia della ricerca. In: Materiali di lavoro. Rivista di studi storici. nuova serie, vol. (1-2), pp. 151–156 (1987)
15. Mordenti, R.: Informatica e critica dei testi. Bulzoni, Roma (2001)

16. Mordenti, R.: Pubblicazione delle ricerche umanistiche in ambiente digitale. In: Archivi informatici per il patrimonio culturale, Novembre, 17-19, 2003. Convegno Internazionale organizzato in collaborazione con ERPANET e la Fondazione 'Ezio Franceschini', pp. 95–117. Accademia Nazionale dei Lincei-Bardi, Roma (2006)
17. OECD, Reccommendation of the Council for Enhanced Access and more effective use of Public Sector information (2008), http://www.oecd.org/internet/interneteconomy/40826024.pdf
18. O'Rieger, O.: Framing digital humanities: The role of new media in humanities scholarship. First Monday 15(10), 4 (2010), http://www.firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3198
19. Orlandi, T.: Informatica testuale. Laterza, Teoria e prassi, Bari (2010)
20. Perrault, A.: Digital Dilemmas: The Transformation of Scholarly Discourse in the Humanities. International Journal of the Humanities 2(2), 1755–1761 (2006), http://works.bepress.com/anna_perrault/22
21. Reale, E.: La valutazione della ricerca pubblica: una analisi della valutazione triennale della ricerca. Franco Angeli, Milano (2008)
22. Robinson, P.: Current issues in making digital editions of medieval text – or, do electronic scholarly editions have a future? Digital Medievalist (2005), http://www.digitalmedievalist.org/journal/1.1/robinson/
23. Roncaglia, G.: Scritture digitali. Lettera Internazionale (98), 48–51 (2008)
24. Tammaro, A.M.: Qualità della comunicazione scientifica. Parte 1 e 2. Biblioteche Oggi (7), 104; (8), 74 (2001)
25. Tammaro, A.M.: Scholarly Communication and Academic Presses. FUP, Firenze (2002), http://epress.unifi.it
26. Torres-Salinas, D., Moed, H.F.: Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. Journal of Informetrics 3(1), 9–26 (2009)
27. Vanhoutte, E.: Three barriers to the development of digital tools in and for the humanities (2006), http://www.edwardvanhoutte.org/pub/2006/pptools160606.htm (retrieved March 11, 2012)
28. White, H.D., Boell, S.K., Yu, H., Davis, M., Wilson, C.S., Cole, F.T.H.: Libcitations: A Measure for Comparative Assessment of Book Publications in the Humanities and Social Sciences. Journal of the American Society for Information Science and Technology 60(6), 1083–1096 (2009)

The Evaluation Approach of IPSA@CULTURA

Maristella Agosti[1], Marta Manfioletti[1], Nicola Orio[2], Chiara Ponchia[2], and Gianmaria Silvello[1]

[1] Department of Information Engineering, University of Padua, Italy
{agosti,manfioletti,silvello}@dei.unipd.it
[2] Department of Cultural Heritage, University of Padua, Italy
nicola.orio@unipd.it, ponchiachiara1@gmail.com

Abstract. This paper reports on the original approach envisaged for the evaluation of a digital archive accessible through a Web application, in its transition from an isolated archive to an archive fully immersed in a new adaptive environment.

Keywords: CULTURA project, case study, IPSA digital collection, digital cultural heritage, different categories of users.

1 Introduction

A multidisciplinary team of the University of Padua, constituted by researchers in both Computer Science and Humanities, is currently involved in the European project CULTURA[1], a STREP project that aims to involve and engage new user categories in Humanities digital collections through the development of new adaptive tools.

The CULTURA project started in February 2011. In the first year efforts were mainly directed towards figuring out the best way of opening up a database developed specifically with highly specialized users in mind to new kinds of users with different levels of knowledge and interest in the collection. As a case study we used IPSA (*Imaginum Patavinae Scientiae Archivum*)[2], an online digital archive of illuminated manuscripts created during a previous collaboration between the Department of Information Engineering and the Department for the History of Visual Arts and Music of the University of Padua [1]. The IPSA digital archive is accessible through a web application.

2 Evaluation Approach

Right from the first year of the CULTURA project IPSA was evaluated by both professional researchers and students. Professional researchers were shown the system first and asked to interact freely with it, after which an interview was held

[1] http://www.cultura-strep.eu/
[2] http://ipsa.dei.unipd.it/

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 147–152, 2014.

addressing their particular research needs, wishes and preferences [2]. Evaluations with students were conducted in November and December 2011 with two different cohorts of students – both undergraduates and postgraduates – who were involved in two different trials. The students were asked to interact with the system in different ways for almost an hour. Afterwards, they were asked to give their feedback by filling in an online questionnaire specifically prepared by a team of psychologists from the University of Graz, a partner of the CULTURA project. The results of these evaluations were reported in [3].

In the subsequent months other interactions were carried out, of various types and with different kinds of users, following a schema of loop-interaction, structured in a first phase of eliciting user requirements, then modifying the IPSA web application accordingly and subsequently evaluating the modifications made. Firstly, a re-engineered version of the IPSA web application, improved according to the results of the November and December evaluations, was presented to a different cohort of students. Then IPSA was modified according to the feedback received from this evaluation and afterwards presented to a small group of professional researchers, involving both researchers already familiar with the system and researchers who interacted with IPSA for the very first time.

Recently, part of the IPSA metadata has been integrated in the CULTURA environment, which offers users tools and functions different from those available in the IPSA web application. In December 2012 IPSA within CULTURA was evaluated by a new group of students. Similarly to the November and December 2011 evaluations these students were asked to interact with the system for a reasonable amount of time, and then they were asked to fill in an online questionnaire. The analysis of their feedback is still ongoing. Furthermore, we are also planning to extend the user categories involved in the evaluation of IPSA within CULTURA, and a number of trials with interested people belonging to the general public will be held during the last year of the project. Then it will be possible to compare user perceptions of a specialized digital archive modified to be suitable for new user categories and a system developed from the very beginning to address all kinds of users.

3 Trial with Students

The IPSA trial with students was conducted in April 2012. It was developed specifically for a cohort of 25 postgraduate students in Management of Archival and Bibliographic Heritage and in Modern Languages, who were attending the course on Databases and Internet. We chose this sample because we thought that users with different fields of expertise are more likely to focus their attention on details that may not have been considered by Computer Scientists or Art Historians.

The trial was divided in two parts, which took place within two weeks (2–16 April 2012). Students were asked to carry out a series of tasks which required them interact with the main functions of IPSA, such as establishing links between two images considered to be related in different ways (described as: copied in, not related to,

same tradition of, sibling of, similar to). Notwithstanding the short timespan in which the two parts of the trial were carried out, we were able to collect students' feedback from the first part and immediately insert the requested changes in the system, thus making them ready for the second part of the evaluation. In this way, the users still remembered the previous issues of the system and were better able to assess the solutions proposed. For example, one of the issues in the first part of the trial was establishing a link between two images. In the previous version of the system, the creation of a link was barely intuitive. The user began with the image of interest and did a search to find a second image of interest. During the entire process, a box with the status of the operation was visible at the top of the page. Since finding the second image may require several searches and a certain amount of time, in the meantime users could forget with which image they had started. Following the observations of the students, this box was enlarged and (what is most useful) now includes the thumbnails of the selected starting-image, some helpful text and large, self-explanatory buttons for completing the link or deleting the operation. The new way of creating a link was presented in the second part of the trial and it received positive feedback. Furthermore, students felt that their suggestions were effectively taken into account, and were more motivated to carry out the second part of the trial.

One change in the interface between the first and the second trials involved the insertion of a drop-down menu at the bottom of the wall of images that allows for an intuitive search through all the illuminations contained in the manuscripts of interest.

Another relevant change that we introduced between the two parts of the trial was the normalization of the plant names. The IPSA collection refers to old manuscripts and it is well known that in earlier times botanical names had no standard form. Moreover, the manuscripts held within the archive are written in different languages (Latin, Italian, Venetian dialect etc.), so it is not uncommon to find a number of variations in the name of the same plant. This is problematic, since such a variety entails difficulties in making consistent queries. Two major problems were addressed: spelling and lexical issues. In terms of spelling, graphic variants need to be aligned, i.e. what we call today *absinth* in English is found in IPSA as *absinthium*, *absenço*, *abscinthium*, and there is no explicit link between all these variants, even though they look similar to the human eye. By contrast lexical issues are even trickier than spelling issues as they need philological and linguistic research in order to be solved. For example, consider the plant today known in English as *cucumber*: within IPSA this plant is called *citrollo* in one manuscript and *cogombaro* in another. Human intervention is clearly needed to establish an explicit link between these two names, allowing users to find both in a single query. During the first evaluation session students used a system in which plant names were not normalized. This means they were unable to find images of the same plant within different manuscripts; in contrast, in the second session they experienced a system in which names had been standardized; in fact the original name was maintained and useful metadata were added to also keep track of the other name variants present in the archive. In this way students were able to appreciate the importance of having a standard nomenclature, which allows consistent and functional queries to be made.

The feedback received from the trial clearly shows that the desire for a simplified way to set the link between two images reveals the need of nonprofessional users to be guided through such a specialist collection. This also emerged from the questionnaires they filled in.

Another aspect to consider is that lack of confidence towards the collection also leads to the desire for a more collaborative environment in which users can share their opinions and reflections and benefit from expert users' help. Students would also like to be able to open their research from the collection to the web, in order to easily get more information for the purposes of their research.

As we will see in the next section, many of the desires expressed by students match professional researchers needs for different reasons. Section 5 instead shows how the CULTURA environment addresses these requirements.

4 Interaction with Professional Researchers

In the same way that different student groups were chosen for the evaluation, it was decided for the interaction with professional users to involve three researchers in History of Art specialized in different domains. Because IPSA was created expressly for specialists in History of Illumination, two of the professional users chosen for the interview were scholars expert in this research area, while the third was a researcher in History of Medieval Painting. The interviews took place in May 2012, so slightly after the evaluations with students, and were carried out on a group basis. The interaction was preceded by a short introduction, which aimed to present the results of the work done until then, especially the trial with the students. Particular emphasis was given to the fact that the students' suggestions had been immediately taken into account and that IPSA had been subsequently modified according to their feedback, to make the three interviewees aware of the importance of their opinions and suggestions and to spur open discussion. The new image search procedure in a manuscript catalogue file was introduced and how it works was explained: the user selects the image from the drop-down list – e.g. f. 11v, *nux muscata* – and then a wall of some 30 images which includes the searched-for image is shown an image wall containing the searched-for image is shown. The three interviewees were delighted with this function, but observed that the searched-for illumination – in this case, the *nux muscata* image – should be underlined in the results to make it easier to find, or it should be the first of the image wall. They also suggested improving the image wall by making it numerically smaller (approximately 20 images) with larger images. Indeed, images are the main research subject of Art Historians, which is why the main concern of experts in this field is to have a set of tools available that allow an in-depth analysis of images. The three interviewees were also shown another improvement made after the evaluation with the students: the new way of setting a link between two illuminations. The attention of the professional researchers was deeply focused on the image also in this case, and they asked to be able to zoom in on the illuminations while setting the link. This highlights once again professional user's need of image investigation tools.

One important issue which arose during the discussion was the need for a more collaborative environment. In particular, professional researchers would like to be

able to create one or more personal folders, to save the results of their queries and searches and to decide which results can be seen by other users and by which user. Such a tool could be used by different research groups, and for teaching purposes as well. It can be seen how this particular element matches the students' need for a collaborative environment where they can be helped by experts in the domain. One useful teaching tool would be the possibility of bookmarking images, so that lecturers could bookmark the images shown in their lectures and students could then find them easily in the digital archive.

Professional researchers also expressed the desire for links to other websites, in order to offer the user quick access to different resources that could improve their research, another requirement that matches students' needs.

5 Evaluation of IPSA within the CULTURA Environment

Between May and October 2012 a subset of metadata from IPSA was imported into the CULTURA environment for use as a case study to test the new environment and its functions. The first evaluation of the new system was carried out on 11 December with a group of postgraduate students both of Linguistics and Communications Theories. Students were asked to interact with the system for approximately one hour, accomplishing some easy tasks that made them interact with most of the CULTURA tools: advanced search, annotations, bookmarks, two different kinds of visualization (called "wheel" and "octopus", see Fig. 1) that show the connections between the image of interest and other elements held in the database, and links to other websites

Fig. 1. Wheel visualization in the CULTURA Environment

(Wikipedia, Google, Bing). The task was structured as follows. Firstly, each student was assigned a plant name and had to search for the image of that plant, and then annotate it. Annotations can be made in two different ways: by creating a note or by directly annotating a portion of the illumination. Students were asked to use both the tools, and possibly to bookmark the image. Afterwards, they had to choose one of the two possible visualizations which allow the user to see all the different relations that the image has in the database, e.g. in which manuscript it is held, or who the illuminator that painted it is. In this case, students were asked to find other manuscripts in which the same plant is represented. By doing so, they found at least another illumination of the same plant, which they were asked to annotate.

6 Conclusions

As the new evaluation trials are still an on-going process, it is not currently possible to analyse users' feedback about the CULTURA environment. Nonetheless, some considerations can still be made.

First of all, it must be considered that the CULTURA environment offers users a wide range of tools that provide an easy and engaging approach to the collection and that address the users requirements elicited in the interactions described above. For example, professional researchers, who are mainly interested in analysing the images, are enabled to annotate images of interest in two different ways: by creating a note or by directly annotating a part of the image, e.g. a detail painted with great skill.

The user can choose to keep the annotation private or to make it public. Clearly this matches both students' and professional researchers' needs: a professor in History of Illumination can annotate images with his observations and thoughts, and annotations can be an effective tool to involve students in the research process of professional users', making it extremely clear to them how an expert in the field proceeds in approaching the images.

Acknowledgments. The work reported has been partially supported by the CULTURA project (reference: 269973) within the Seventh Framework Programme of the European Commission, Area "Digital Libraries and Digital Preservation".

References

1. Agosti, M., Benfante, L., Orio, N.: IPSA: A Digital Archive of Herbals to Support Scientific Research. In: Sembok, T.M.T., Zaman, H.B., Chen, H., Urs, S.R., Myaeng, S.-H. (eds.) ICADL 2003. LNCS, vol. 2911, pp. 253–264. Springer, Heidelberg (2003)
2. Agosti, M., Orio, N.: User Requirements for Effective Access to Digital Archive of Manuscripts. Journal of Multimedia 7(2), 217–222 (2012)
3. Ponchia, C.: Engaging the User: Elaboration and Execution of Trials with a Database of Illuminated Images. In: Agosti, M., Esposito, F., Ferilli, S., Ferro, N. (eds.) IRCDL 2012. CCIS, vol. 354, pp. 207–215. Springer, Heidelberg (2013)

Optimizing Relevance Ranking to Enhance the User's Discovery Experience*

Tamar Sadeh

Ex Libris Group, Jerusalem, Israel
tamar.sadeh@exlibrisgroup.com

Abstract. With the introduction of library discovery systems, the display of results according to relevance, as determined by the system, has become a norm. To investigate how relevance ranking could be optimized, a provider of a widely used discovery system developed methods of evaluating the system's relevance ranking. As a result, new factors were added to the calculation of search results' relevance—information about the individual user and the user's information needs, and an indicator representing the academic significance of materials. Methods of monitoring the impact of changes were also established.

Keywords: relevance ranking, user experience, discovery systems, personalized ranking.

1 Introduction

Developed at the turn of the millennium to enable users with little information literacy to reach general information without the help of mediators (such as travel agents, salespeople, or librarians), Web search engines have shaped the way in which students and researchers seek scholarly information today. However, the ease of use, immediacy of results, and heterogeneous nature of information provided by Web search engines such as Google trigger expectations that libraries have only recently started to meet with "new-generation" library discovery systems (e.g., [1], [2], [3]).

These index-based discovery systems, available from 2009, aim to provide a single entry point to the scholarly information landscape and enable users to search in a Google-like way in local, global, and regional collections that libraries offer their users. Such global and regional collections, provided by primary and secondary publishers and aggregators, include journal articles, e-books, conference proceedings, newspaper articles, theses, patents, and other types of materials. The collections differ in many ways, such as in the type of content (metadata, abstract, or full text), the format and depth of the available metadata, and the licensing options.

Although new-generation discovery systems may match Web search engines in ease of use and speed, the success of a search process is measured not just by the amount of time that elapses until a result list is displayed. Much more important is the

* To protect Ex Libris Ltd.'s intellectual property, some details are not included in this paper.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 153–164, 2014.

amount of time that an information system takes to satisfy a user's need and furnish the desired outcome. A key challenge of library discovery systems is how to provide users with the most relevant items from the immense landscape of available content. To meet this challenge, developers have enriched systems with new features. For example, faceted navigation helps users quickly refine their result list and focus on its subsets (see [4] and [5]), and recommendations based on other users' prior selections draws searchers' attention to items related to the topic of their search, even when such articles do not match the query terms that the searchers entered.

However, because of their familiarity with Google and other search engines, users of discovery systems tend to scan only the topmost results; hence, items that are most relevant for a particular search can easily remain unnoticed if they are not displayed near the top of the list. Whereas past discussions of relevance in information retrieval distinguished between what is relevant and what is not, the current focus is on the *degree* of relevance, with the understanding that "in the most fundamental sense, relevance has to do with effectiveness of communication" [6]. Relevance ranking, whose purpose is to highlight materials that the system deems the most pertinent for the particular query, has become a major factor in satisfying user needs. Together with the immediate delivery of retrieved items, relevance ranking has had a huge impact on increasing the value of library services for users and institutions.

2 Relevance in the Scholarly Domain

Although "relevance is a, if not even *the*, key notion in information science in general and information retrieval in particular" [7], the application of relevance in library information systems developed in the past was (and in some of those systems, continues to be) limited to a binary approach: a document was considered either relevant or not relevant to a specific query. The determination was based on the textual similarity between the item and the query; the user's specific information need, among other things, was not considered, and results were not arranged according to the degree of relevance to the specific user. Rather, results were displayed according to unambiguous criteria, such as the date or the alphabetic order of the title or author name.

Over the past decade, library information systems entered a new realm and adopted relevance ranking so that result lists would address user expectations. Harnessing traditional relevance-ranking methods used by other information systems, library systems adjusted the methods for scholarly materials.

The theoretical aspects of relevance ranking have been studied for many years [e.g., 8], but literature on practical implementations has been lacking. In an attempt to develop an automated calculation of relevance that includes information about the user and the user's information need and leverages well-structured metadata and other data available about scholarly materials, one discovery system vendor, Ex Libris, conducted a research program whose results transformed the discovery system's relevance ranking.

3 Direct and Inferred Assessment of Relevance

One of the major challenges in developing relevance ranking is assessing its success. Because relevance is subjective, a qualitative evaluation of success depends on the characteristics of each individual who tests the information system. One's specific information needs, expectations, and expertise in both the area of research and searching techniques come into play when one assesses the relevance of displayed results.

Another approach, less dependent on the engagement of individuals in testing, is to compare the order in which a system displays the results with the order of results in another system that is known to have excellent relevance ranking. This approach, while easy to implement, has major shortcomings. First, no scholarly information system has a relevance-ranking method that is reliable enough to serve as a model. Second, the content available in every such system is different: in a discovery system, the content depends on the institution's subscriptions and policies; when it comes to Google Scholar or Microsoft Academic Search, the exact search scope is not documented. Third, the format, degree of completeness, and depth of the available metadata (basic or comprehensive metadata) and the presence or absence of full text in addition to metadata differ from one system to another.

An alternative approach is to quantitatively test how modifications in relevance-ranking technology affect searchers' behavior. An analysis of usage data from a large number of users can shed light on their overall satisfaction; measures such as an increase in the number of sessions that culminate in the selection of an item and a higher average position of selected items in the result list are likely to indicate that the relevance ranking has improved.

4 Relevance-Ranking Project

4.1 Aim and Objectives of the Project

In 2010, a centralized, cloud-based index of hundreds of millions of global and regional scholarly materials such as journal articles, e-books, conference proceedings, and newspaper articles was added to the Ex Libris Primo® discovery system. As a result, users at Primo libraries can launch a single search that covers both their local library holdings and an information landscape that transcends the traditional boundaries of the library's offering. This landscape includes subscribed and open-access materials and, if the library so desires, materials that the library or the user can purchase on demand. Because of the amount of data available, the heterogeneous nature of this virtual collection, and users' tendency to implement "simple" search queries [9], [10], the Primo system's relevance ranking had to be enhanced to effectively support the needs and expectations of the users.

To investigate the ways in which Primo relevance ranking should be improved, Ex Libris assembled a team of information retrieval experts, research and development staff, and product managers. In March 2011, the team began a long-term project

to support the continued development of the Primo relevance-ranking technology. The team defined the following project objectives:

- Examine search logs and usage statistics to understand search trends and the ways in which Primo accommodates the various search practices of end users
- Set a baseline for the assessment of Primo relevance ranking by obtaining researchers' evaluations of the current sorting order of the results
- Define metrics to evaluate the effectiveness of Primo relevance ranking
- Build a test environment in which the defined metrics are used to test and monitor changes to the technology
- Optimize the calculations underlying Primo relevance ranking and incorporate additional information about an item (the item's "value score"), the user, and the user's needs at the time of the search
- Monitor real-world Primo users' search behavior over time to evaluate the impact of the improvements and suggest further improvements

4.2 Test Environment and Testing Tools

To set a baseline that would later serve as a reference point and to test the changes that would be introduced to the relevance-ranking technology, the project team set up a test environment at the Ex Libris research and development lab and developed testing tools. The methods for evaluating relevance ranking in the test environment were based on both a qualitative evaluation by users and a comparison with Google Scholar.[1] This comparison was not limited to an examination of the differences between the result lists of Primo and Google Scholar; rather, the examination included input from researchers regarding both lists.

The test environment consisted of a Primo implementation with a search scope of 40 million records, a representative sample of the content indexed in the Primo Central Index at the time of the testing. In addition, the team developed two testing tools:

- A tool that makes ad hoc changes to parameters affecting the ranking method but does not require reindexing of the data or changes in the software. With this tool, the team could further boost materials published in the current year.
- A tool that runs test queries and calculates two common measures of quality[2]:
  - Mean average precision (MAP), which indicates how successful the system was in ranking "best" items above other items. A "best" item is one that should appear on the first page of results.
  - Mean reciprocal rank (MRR), which indicates how high in the result list the first "best" item appears

[1] Both Primo and Google Scholar index a large number of academic publications of various types and from many information providers, and enable users to search via a single, simple interface. Despite differences between these systems (primarily, libraries' control over the search scope, user interface, and integrated services), the search experience is similar.

[2] See `http://en.wikipedia.org/wiki/Mean_average_precision` and `http://en.wikipedia.org/wiki/Mean_reciprocal_rank`.

MAP and MRR metrics typically distinguish between relevant items and non-relevant items in an entire result list. However, for the purposes of the project's relevance-ranking tests, a distinction was made between "best" results—those that should appear on the first page—and "other" results (which may be relevant to the query but are not necessarily relevant enough to be on the first page). Furthermore, the MAP and MRR values were calculated for only the top 20 results.

4.3 Setting a Baseline

The project began with a user evaluation of the Primo relevance ranking that was available in March 2011. To enable the team to obtain a more objective assessment, the evaluation included a comparison between Primo relevance ranking and that of Google Scholar (the team kept in mind the caveats regarding such a comparison, as previously mentioned). The data from the evaluation formed a baseline to help the team assess the impact of future improvements.

A group of senior researchers from several institutions agreed to evaluate the results of the Primo relevance ranking. These evaluators work in various disciplines: agriculture, anthropology, biology, medicine, military and security studies, philosophy, and physics.[3] Each evaluator created search queries related to his or her area of expertise, including broad-topic queries, narrow-topic queries, known-item queries, and other types (Fig. 1). In addition, the project team conducted in-depth interviews with three evaluators to learn what they expect from a scholarly information system.

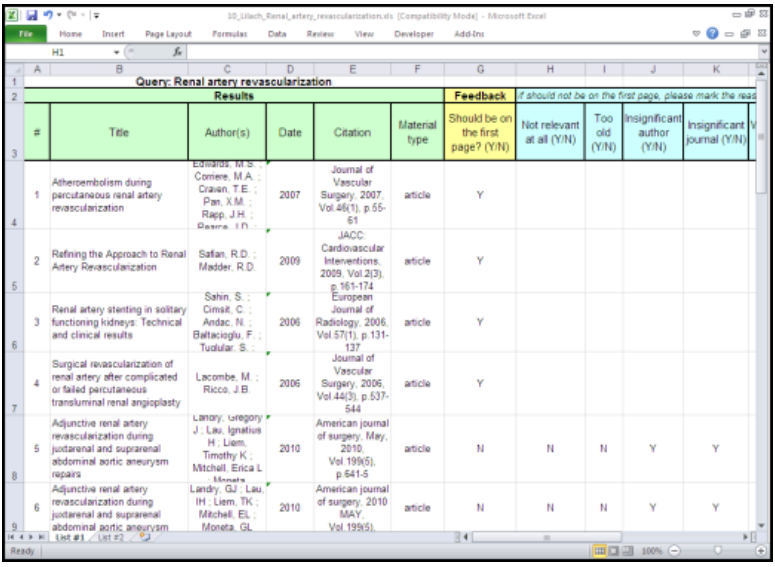
Flavor in soft leptogenesis
gated super conductivity
halperin lee read
labor process & structures of feeling
Particle cosmology
marxist anthropology
anthropology of class israel
ranciere nights of labor
chloroplast
photoinhibition
the role of proteases in repair from photoinhibition
Twin earth thought experiment
Frank Jackson what mary didn't know the journal of philosophy
Renovascular hypertension
Controversies in renal artery stenosis: a review by the American Society of Nephrology Advisory Group on Hypertension

Fig. 1. A sample of the experts' queries

The team ran the queries in Primo and Google Scholar and sent each evaluator a spreadsheet that listed, for each query, the first 20 results returned by Primo and the first 20 results returned by Google Scholar. To exclude bias as a factor in the evaluation, the team did not inform the evaluators of the results' origin. The evaluators' task was to indicate whether each item should be displayed on the first page of results. Also, for each item that the evaluators said should not be on the first page, they were

[3] The decision to set a baseline with the help of experts was based on the assumption that these experts know which results to expect when submitting their queries.

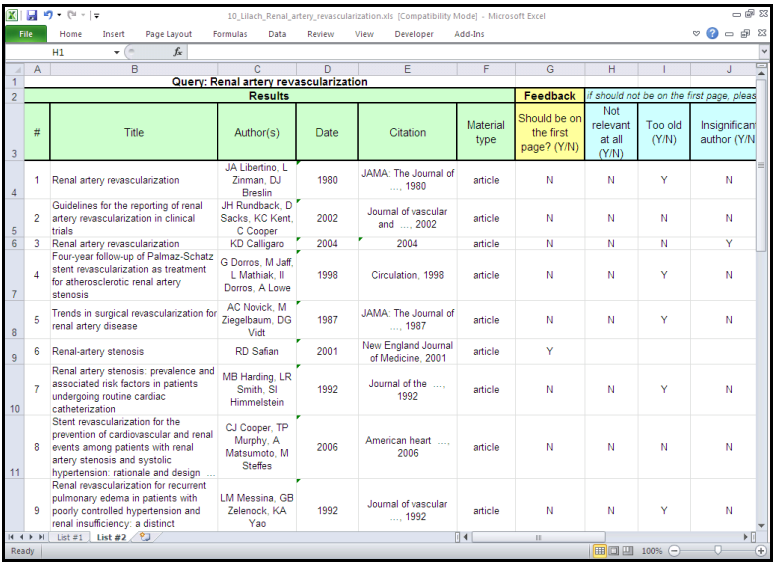
asked to explain why, by selecting at least one reason from a list or writing their own reasons (Fig. 2 and Fig. 3). The spreadsheet offered the following reasons: not relevant at all, too old, insignificant author, insignificant journal, wrong material type, too specific, too broad, and not the specific article the researcher was looking for.



Query: Renal artery revascularization

		Results					Feedback	If should not be on the first page, please mark the reas				
	#	Title	Author(s)	Date	Citation	Material type	Should be on the first page? (Y/N)	Not relevant at all (Y/N)	Too old (Y/N)	Insignificant author (Y/N)	Insignificant journal (Y/N)	V
4	1	Atheroembolism during percutaneous renal artery revascularization	Edwards, M.S. ; Comiere, M.A. ; Craven, T.E. ; Pan, X.M. ; Rapp, J.H. ; Reeves, J.D.	2007	Journal of Vascular Surgery, 2007, Vol.46(1), p.55-61	article	Y					
5	2	Refining the Approach to Renal Artery Revascularization	Safian, R.D. ; Madder, R.D.	2009	JACC: Cardiovascular Interventions, 2009, Vol.2(3), p.161-174	article	Y					
6	3	Renal artery stenting in solitary functioning kidneys: Technical and clinical results	Sahin, S. ; Cimsit, C. ; Andac, N. ; Baltacioglu, F. ; Tuzular, S. ;	2006	European Journal of Radiology, 2006, Vol.57(1), p.131-137	article	Y					
7	4	Surgical revascularization of renal artery after complicated or failed percutaneous transluminal renal angioplasty	Lacombe, M. ; Ricco, J.B.	2006	Journal of Vascular Surgery, 2006, Vol.44(3), p.537-544	article	Y					
8	5	Adjunctive renal artery revascularization during juxtarenal and suprarenal abdominal aortic aneurysm repairs	Landry, Gregory J ; Lau, Ignatius H ; Liem, Timothy K ; Mitchell, Erica L ; Moneta,	2010	American journal of surgery, May, 2010, Vol.199(5), p.641-5	article	N	N	N	Y	Y	
9	6	Adjunctive renal artery revascularization during juxtarenal and suprarenal abdominal aortic aneurysm	Landry, GJ ; Lau, IH ; Liem, TK ; Mitchell, EL ; Moneta, GL	2010	American journal of surgery, 2010 MAY, Vol.199(5),	article	N	N	N	Y	Y	

Fig. 2. Results returned by Primo and evaluated by a medical researcher



Query: Renal artery revascularization

		Results					Feedback	If should not be on the first page, pleas			
	#	Title	Author(s)	Date	Citation	Material type	Should be on the first page? (Y/N)	Not relevant at all (Y/N)	Too old (Y/N)	Insignificant author (Y/N	
4	1	Renal artery revascularization	JA Libertino, L Zinman, DJ Breslin	1980	JAMA: The Journal of ..., 1980	article	N	N	Y	N	
5	2	Guidelines for the reporting of renal artery revascularization in clinical trials	JH Rundback, D Sacks, KC Kent, C Cooper	2002	Journal of vascular and ..., 2002	article	N	N	N	N	
6	3	Renal artery revascularization	KD Calligaro	2004	2004	article	N	N	N	Y	
7	4	Four-year follow-up of Palmaz-Schatz stent revascularization as treatment for atherosclerotic renal artery stenosis	G Dorros, M Jaff, L Mathiak, II Dorros, A Lowe	1998	Circulation, 1998	article	N	N	Y	N	
8	5	Trends in surgical revascularization for renal artery disease	AC Novick, M Ziegelbaum, DG Vidt	1987	JAMA: The Journal of ..., 1987	article	N	N	Y	N	
9	6	Renal-artery stenosis	RD Safian	2001	New England Journal of Medicine, 2001	article	Y				
10	7	Renal artery stenosis: prevalence and associated risk factors in patients undergoing routine cardiac catheterization	MB Harding, LR Smith, SI Himmelstein	1992	Journal of the ..., 1992	article	N	N	Y	N	
11	8	Stent revascularization for the prevention of cardiovascular and renal events among patients with renal artery stenosis and systolic hypertension: rationale and design ...	CJ Cooper, TP Murphy, A Matsumoto, M Steffes	2006	American heart ..., 2006	article	N	N	N	N	
	9	Renal revascularization for recurrent pulmonary edema in patients with poorly controlled hypertension and renal insufficiency: a distinct	LM Messina, GB Zelenock, KA Yao	1992	Journal of vascular ..., 1992	article	N	N	Y	N	

Fig. 3. Results returned by Google Scholar and evaluated by the same medical researcher who evaluated the results in Fig. 2

The project team received the evaluations and calculated the lists' MAP and MRR. Despite the fact that the test environment covered only 40 million Primo Central Index records, the results clearly showed that as a starting point for the project, the Primo ranking was respectable. For narrow-topic searches, the MAP of Primo was better than that of Google Scholar; for known-item searches, the Primo MAP was surprisingly high in light of the fact that the data searched was much smaller than Google Scholar's; and for broad-topic searches and "other" searches (not one of these three types), Google Scholar's MAP was better (Table 1).

Table 1. Baseline mean average precision scores for Primo and Google Scholar results

Query type	Primo MAP	Google Scholar MAP
Broad-topic	0.31	0.72
Narrow-topic	0.52	0.37
Known-item	0.52	0.67
Other	0.36	0.70

An analysis of evaluators' reasons for disqualifying results from inclusion on the first result page showed that in 29% of the cases, the journal in which the paper was published was deemed insignificant; in 20% of the cases, the evaluator considered the author insignificant; and in 15% of the cases, the material (typically an article) was too specific. These reasons do not apply to known-item searches, in which users want a specific item, regardless of the prominence of the author or journal. In exploratory searches, which can yield a very large number of results, evaluators expected results of a general nature, such as the most important review articles on a topic from top journals and leading researchers. Evaluators disqualified 7% of the items for being too old.

The queries from the initial tests were stored with the researchers' evaluations and later served as queries for testing new enhancements to the Primo relevance ranking.

4.4 Optimizing the Basic Relevance-Ranking Technology

Before introducing new factors, the project team set out to improve the relevance-ranking technology available in March 2011. This technology used traditional ranking methods that had been adapted to the scholarly domain. One method, for example, was the weighting of metadata fields according to the significance of the bibliographic detail stored in them; hence, query terms in a subject field, for instance, contributed more value to a specific item's ranking than the terms' occurrence in the full text.

On the basis of the evaluators' input and customer feedback on the system, the team leveraged the built-in boosting mechanism of Primo: using a tool developed for the testing, the team changed the boosting factors and then monitored the impact. For example, although journal articles were already more prominent than other material types, the team decided to boost these articles even more. The resulting MAP and MRR values peaked when the boosting factor was roughly 1.5 (Fig. 4), indicating that 1.5 is the optimal boosting factor for this material type.

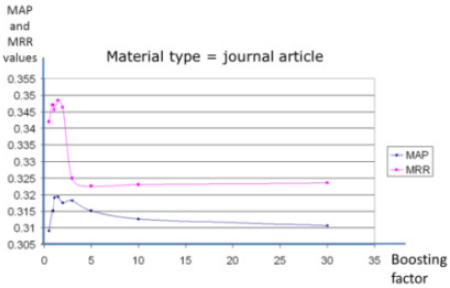


Fig. 4. MAP and MRR values after each change in the boosting factor for the journal-article material type

4.5 Adding a Value Score

After optimizing the technology, the team decided to introduce the first new factor. As a result of the interviews and the evaluators' responses, the team realized that the scholarly significance of an item should play a role in determining its position in the result list. This new factor, the item's *value score,* can be compared to Google PageRank in that both are attributes of an "item" (a Web site in Google searches) and are independent of a given query.

The team determined that value scores would be based on various types of information, including the number of citations and usage data from the Ex Libris bX article recommender database.[4] After incorporating value scores in the relevance ranking, the team recalculated the MAP and MRR values for the test queries and found that, indeed, the values went up, indicating that the addition of a value score improved the relevance ranking and that the "best" items were now located farther up the result list.

In the future, the team plans to enrich value scores with other types of information, such as the scholarly impact of a journal or an author.

4.6 Monitoring the Changes

Primo 3.1, released in June 2011, incorporated these changes in the relevance-ranking method. The project team monitored the impact of the changes on users' information-seeking patterns, examining the following factors in particular:

- The mean number of times per session that a user moved to the next page of results. The team assumed that users would typically go to the next page only when their information need was not satisfied by the items already displayed.
- The percentage of sessions that culminated in the selection of an item[5]
- The mean number of seconds that elapsed from the beginning of a session until an item was selected
- The mean position of the selected items in the result list

[4] See `http://www.exlibrisgroup.com/category/bXRecommender`.
[5] A selection, in this context, is the user's explicit request to obtain an item.

4.7 Gathering More Data

Before moving on to further improvements, the project team extended the corpus of test queries by engaging more users—graduate students recruited from several universities and various academic fields. This time, the test environment enabled the evaluators to access the hundreds of millions of records covered by the Primo Central Index (as opposed to the 40 million records available for the first evaluators). In addition, the Primo interface in the test environment was enhanced with thumbs-up and thumbs-down buttons next to each of the first 20 results (Fig. 5). The testers used the buttons to indicate whether an item should or should not be located on the first result page. The system logged each query, its first 20 results, and the users' feedback.

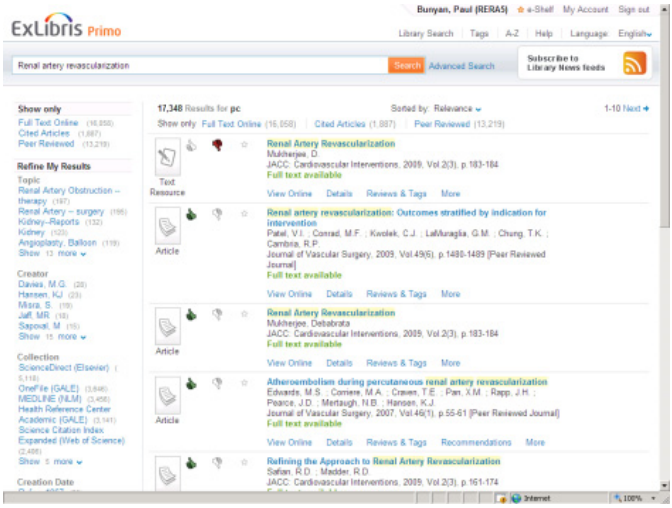


Fig. 5. New test environment for providing feedback on the relevance of materials displayed on the first page of results

This test ran for nine weeks beginning in August 2011, and 560 valid queries with evaluations were logged. The mean MAP and MRR values of all the queries (0.69 and 0.79, respectively) were much higher than the values from the initial testing phase (Fig. 6). The improvement was attributed primarily to the relevance-ranking technology's enhancements and the availability of the full scope of the index for searching. However, the fact that the evaluators were graduate students rather than senior researchers is likely to have played a role as well: graduate students' information needs probably differ from those of senior researchers, and the students are perhaps less familiar with the available content than more experienced researchers are. In addition, graduate students might be less confident in their understanding of what should and should not be on the first page of results. Indeed, Anderson points out that "the evaluation of relevance is thereby linked to the measure of uncertainty…Previous knowledge (or lack of knowledge) is indispensable in assessments of relevance" [11].

121	CO2 dynamics	0.816667	1	Graduate	Physics, Astronomy, ar
122	silicon fractionation	0.494946	0.5	Graduate	Physics, Astronomy, ar
123	silicon isotopes fractionation	0.717248	1	Graduate	Physics, Astronomy, ar
124	phreeqc	0.52243	1	Graduate	Physics, Astronomy, ar
125	quartz kinetics	0.545673	1	Graduate	Physics, Astronomy, ar
126	albite dissolution kinetics	0.889727	1	Graduate	Physics, Astronomy, ar
127	kaolinite precipitation kinetics	0.630305	1	Graduate	Physics, Astronomy, ar
128	Gypsum Nucleation	0.756605	1	Graduate	Geology
129	Gypsum Nucleation Dead sea	0.601651	1	Graduate	Geology
130	Gypsum crystal grows	0.71622	1	Graduate	Geology
131	Gypsum Kinetics	0.633892	0.5	Graduate	Geology
132	red sea dead sea conduit	0.700735	1	Graduate	Geology
133	gypsum preciptation potential	0.833333	1	Graduate	Geology
134	Microbial ecology Dead Sea	1	1	Graduate	Geology
135	rate law for gypsum nucleation	0.868034	1	Graduate	Geology
136	radium	1	1	Graduate	Biology, Life Sciences,
137	radium AND barium	0.67006	1	Graduate	Biology, Life Sciences,
138	radium AND barium co-precipitation	0.721398	1	Graduate	Biology, Life Sciences,
139	activity coefficient Pitzer	0.874559	1	Graduate	Biology, Life Sciences,
140	activity coefficient Pitzer AND radium	1	1	Graduate	Biology, Life Sciences,
141	activity coefficient Pitzer AND barite	0.738095	1	Graduate	Biology, Life Sciences,
142	activity coefficient Pitzer AND gypsum	0.421775	0.5	Graduate	Biology, Life Sciences,
143	solid solution	0.227273	0.25	Graduate	Biology, Life Sciences,
144	solid solution AND barite	0.826356	1	Graduate	Biology, Life Sciences,

Fig. 6. A sample of MAP (second column) and MRR (third column) values calculated for the results displayed on the first page, as evaluated by graduate students

4.8 Adding the User to the Equation

The next phase focused on adjusting the relevance-ranking calculations to suit certain characteristics of the individual and his or her specific information need as evidenced in the query. Primo 4.0 (released in June 2012) incorporated these adjustments.

Broad-Topic Queries. According to an analysis of search logs, 25% to 30% of the queries submitted by Primo users are general in nature or applicable to many areas. Search logs have yielded many examples of broad-topic queries, such as *korean poetry*; *buddhist art*; *adventure education*; *stereoisomerism*; *social mobility*; *mining engineering*; *operator theory*; *sensitivity*; and *tetrodotoxin*.

When users enter a broad-topic query, they do not have a specific document in mind and usually receive a large result set. The project team's assumption was that users invoke a broad-topic query when they want to explore an area with which they are not very familiar—for example, undergraduates who are seeking material for a class assignment about a general topic or researchers who are looking for information related to a discipline with which they are not familiar.[6]

The team also assumed that for broad-topic queries, users would prefer results that provide general information. For example, when a user submits the query *operator theory*, the most appropriate results would be documents that explain what operator theory is, such as reference materials or review articles, rather than articles addressing specific issues related to operator theory.

To improve the ranking of results from broad-topic queries, the team made several adaptations to the relevance-ranking calculations, including offline processing, ways of identifying broad-topic queries in real time, and the boosting of materials that are more likely to address the specific needs of users who formulate this type of query.

[6] According to a survey by University of Minnesota Libraries, the types of queries that undergraduates submit differ significantly from queries that graduate students submit: over 70% of undergraduate searches are exploratory, but as researchers become more knowledgeable in their area, they tend to search more for specific items [12].

Author-Related Queries. According to search logs, about 10% of Primo queries consist of author names (sometimes with other words) in the default search box (as opposed to names that were typed in the author field). Search logs have yielded many examples of what appear to be author-related queries, such as *wescott, D and oil and australia*; *Peterson, T (2004)*; *walt whitman leaves of grass*; *"Nam Soon Huh"*; *Text types adults Nunan and Lamb*; and *Nemes & Coss Effective Legal Research*.

As with broad-topic queries, the team adapted the relevance-ranking calculations to improve the handling of author-related queries. These adaptations include offline processing, methods of identifying author names in real time, and the rephrasing of queries to better address the presumed intention of the user.

Personalized Ranking. Personalized ranking shows much promise and can be addressed on several levels. By deploying such ranking, a discovery system may be able to minimize the number of irrelevant results stemming from topic ambiguity; furthermore, the system can adjust the type of materials to the user's level of expertise. In interdisciplinary research specifically, any division of the search scope by formal disciplinary boundaries may prevent successful discovery; however, an all-encompassing system that tailors the results to the particular user will improve the likelihood of a successful outcome.

Today, the Primo technology takes into account a user's academic degree and discipline (without identifying the person). Users decide whether they wish to provide these details, and if so, the system keeps the information in the users' personal profile, where they can modify it.

Identifying a user's academic degree enables Primo to provide the most suitable results for a topic search. Undergraduates probably need general materials, whereas experienced researchers are likely to seek more in-depth publications. Therefore, on the basis of a user's degree, Primo boosts specific material types for topic searches.

With information about a user's discipline, Primo can boost items in that area—a particularly important feature when search terms apply to several areas. For example, unless the user's discipline is factored into the ranking calculation, the query *memory efficiency* yields results in engineering, medicine, psychology, and other areas. Similarly, when an author-related query contains a common name, the relevance-ranking technology can boost items from authors who publish in the user's field.

5 Conclusions

The relevance-ranking project has demonstrated the feasibility of a methodic approach to developing and testing methods of relevance ranking in a scholarly information system. Although the relevance of a document is always a function of a specific user at a certain point in time, some aspects of relevance can be generalized and deployed in an automated system. The project described here shows that when relevance ranking incorporates the scholarly significance of items, assumptions about a user's type of information need as inferred from the query, and a correlation between a user's academic area and the topics of available materials, the probability that the most relevant items will be displayed at the top of the result list increases.

Nevertheless, the calculation of relevance remains an area that requires constant rethinking, monitoring, and tuning. We can only agree with Saracevic's remark that "there were, still are, and always will be many problems with relevance. This is not surprising. Relevance is a human—not a systems—notion and human notions are complex, even messy. Oh well, they are human" [7].

References

1. Centre for Information Behaviour and the Evaluation of Research (CIBER): Information Behaviour of the Researcher of the Future (2008)
2. OCLC Online Computer Library Center: Perceptions of Libraries and Information Resources: A Report to the OCLC Membership. OCLC, Dublin, Ohio (2005)
3. OCLC Online Computer Library Center: College Students' Perceptions of Libraries and Information Resources: A Report to the OCLC Membership. OCLC, Dublin, Ohio (2006)
4. Sadeh, T.: Multiple Dimensions of Search Results. Paper given at Analogous Spaces Interdisciplinary Conference. Ghent University, Belgium (2008)
5. Hearst, M.A.: Clustering versus Faceted Categories for Information Exploration. Communications of the ACM 49(4), 59–61 (2006)
6. Saracevic, T.: Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science. J. Am. Soc. Inf. Sci. Technol. 26(6), 321–343 (1975)
7. Saracevic, T.: Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. J. Am. Soc. Inf. Sci. Technol. 58(13), 1915–1933 (2007)
8. Maglaughlin, K.L., Sonnenwald, D.H.: User Perspectives on Relevance Criteria: A Comparison among Relevant, Partially Relevant, and Not-Relevant Judgments. J. Am. Soc. Inf. Sci. Technol. 53(5), 327–342 (2002)
9. Markey, K.: The Online Library Catalog: Paradise Lost and Paradise Regained? D-Lib Magazine 13(1/2) (2007)
10. Markey, K.: Twenty-five Years of End-User Searching, Part 2: Future Research Directions. J. Am. Soc. Inf. Sci. Technol. 58(8), 1123–1130 (2007)
11. Anderson, T.D.: Uncertainty in Action: Observing Information Seeking within the Creative Processes of Scholarly Research. IR Information Research 12(1) (2006)
12. Hanson, C., Hessel, H., Boudewyns, D., Fransen, J., Friedman-Shedlov, L., Hearn, S., Herther, N., Theis-Mahon, N., Morris, D., Traill, S., West, A.: Discoverability: Phase 2. Final Report, University of Minnesota Libraries (2010)

Sapienza Libraries and Google Books Project

Adriana Magarotto, Maura Quaquarelli, and Mattia Vallania

Sistema Bibliotecario Sapienza, Sapienza - Università di Roma, Italia
{adriana.magarotto,maura.quaquarelli,
mattia.vallania}@uniroma1.it

Abstract. The report shortly examines the experience of Sapienza libraries as partners of Google Books project, signed in July 2011. The goal is to digitize 35,000 books from 1500 to 1872 during the first year of activity. The issue concerns management and the optimization of bibliographic data set, development of web-based instruments for ruling the workflow and sharing records and information between the ILS system (Sebina Open Library) and external data bases.

Keywords: libraries, digitization, organization, catalogues, bibliographic data.

1 Introduction and Contest

A research library's mission is to set up document collections to satisfy user needs, so it's necessary to ensure access to these collections and to make it possible to spread historical memory and knowledge kept in these libraries. Recent digitization projects, many economical and space problems for the printed collections' growing have given new opportunities and changed the concept of research library. To pursue this transformation, a process going on in libraries worldwide, Sapienza decided to take part in the Google Books project, a precious chance to jumpstart complete digitization. Previously other similar projects have been realized by single departments or in partnerships, such as ProDigi (2008-2009). These projects have made it possible to create an archive of digitized texts, and helped spread knowledge and best practices which we can now use for this brand new project, a complex effort to create and increase Sapienza Digital Library.

After the agreement between Sapienza University of Rome and the Ministry of Cultural Heritage-MiBAC, signed by the University Rector in July 2011, Sapienza was willing to give a large part of its own book collections for the digitization. Now the project is running and many books have been sent since November 2011. We think that we can digitize about 35,000 books during the first year. We start with ancient printed books, from 1500 to 1700 up to but not over1872, a conventional date established by international copyright laws. Only 10 libraries have taken part in the first step of the project, because of time constraints and the experimental nature of the project. The Google contract expects maximum privacy on their technical solutions, so the following description briefly shows problems, solutions and results of the first part, only for SBS-Sapienza.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 165–170, 2014.

2 Technical and Organizational Characteristics

In the first operative phase we had to select, prepare and send all those books conform to Google standards. We have tackled some problems such as the format for cataloguing data, books shipment and organization of the work between the SBS Centre and ten libraries.

The main requirements asked by Google to digitize the documents are:

All the books must have metadata. We get the metadata from the collective catalogue of polo RMS (we call polo a local part of our National Library Service, SBN). This catalogue has SOL software, an integrated library system (ILS), realized by Data Management; the software is a web application that manages all the main librarian functions (cataloguing, purchasing, lending, users database management), for back and front office. Most of scientific books and rare editions of Sapienza have already been catalogued. The software makes the data export from SBN format possible in order to exchange bibliographic data formats used both in Europe and USA, Unimarc and Marc21. We send a Marc21/xml file to identify our records in Google Books.

All the books must have a barcode with a univocal identification code, unique for all Sapienza libraries. The barcode reading is necessary in every phase: book shipment, digitization, metadata and book linking that will be available in Google Research Interface.

The Sapienza collection must be considered as one collection, even if we have several collections in different places that belong to economically and organizationally independent centers. The collective catalogue is very useful in this situation, but it gives us solutions just for a part of our problems. The organization of the activities and the communication with Google make it necessary to produce too many files and printed texts.

3 Solutions

First, it's necessary to develop an instrument, an operative context because of two main reasons:

- to limit the costs of developing a new component in SOL, used just in this temporary project
- to have system components which are quickly increased and adapted to our operative needs

For the realization of these instruments there is a team of SBS staff, librarians and software developers of Cineca, Sapienza's partner with a great experience in bibliographic data management and technical manager of RMS polo.

Here we have a brief description of the problems we tackled.

3.1 Data Format

Bibliographic data export to Marc21, generated by SOL software, is transformed in Marc21/xml according to a standard; but some modifications are necessary on metadata, as requested by Google, especially on multi-volume works' titles that must start and end in field 245. So this field has been appropriately modified, using $n and $p subfields for hierarchical links (see the example)

For the administrative section we add the own identification code in field 955.

```
<record>
<leader>00869nam 22001937i 4500</leader>
<controlfield tag="001">PAR0736263</controlfield>
<controlfield tag="008">100224s1869 it ||||  |
|||||||ITAod</controlfield>
<datafield tag="041" ind1="0" ind2=" ">
<subfield code="a">ITA</subfield>
</datafield>
<datafield tag="100" ind1="1" ind2=" ">
<subfield code="a">Curioni, Giovanni</subfield>
</datafield>
<datafield tag="245" ind1="1" ind2="0">
<subfield code="a">L'arte di fabbricare, ossia Corso
completo di istituzioni teorico-pratiche per gli
ingegneri, per gli architetti, pei periti in costruzione
e pei periti misuratori</subfield>
<subfield code="p">Operazioni topografiche</subfield>
<subfield code="c">per Giovanni Curioni</subfield>
</datafield>
<datafield tag="260" ind1=" " ind2=" ">
<subfield code="a">Torino</subfield>
<subfield code="b">A. F. Negro</subfield>
<subfield code="c">1869</subfield>
</datafield>
<datafield tag="300" ind1=" " ind2=" ">
<subfield code="a">351 p.</subfield>
<subfield code="c">25 cm.</subfield>
</datafield>
<datafield tag="774" ind1=" " ind2="0">
<subfield code="t">L'arte di fabbricare, ossia Corso
completo di istituzioni teorico-pratiche per gli
ingegneri, per gli architetti, pei periti in costruzione
e pei periti misuratori</subfield>
<subfield code="w">RMS191070</subfield>
</datafield>
<datafield tag="852" ind1=" " ind2=" ">
```

```
<subfield code="a">RMSAR</subfield>
<subfield code="c">ARlibro MINN. C 838 </subfield>
<subfield code="t">AR 33621 </subfield>
</datafield>
<datafield tag="955" ind1=" " ind2=" ">
<subfield code="a">BIBLIOTECA CENTRALE DELLA FACOLTA' DI
ARCHITETTURA</subfield>
<subfield code="z">RMSAR$$$000033621$$$D</subfield>
</datafield>
</record>
```

[Example of a modified marc21/xml file]

3.2 Barcode

The barcode is realized according to standard "code 39", it's made of 20 symbols (letters or numbers) and a check digit. Every barcode starts with RMS, the library system name; then we have 2 letters that identify the specific library and 15 symbols that represent the inventory number linked with the book. We use $ when we have an empty space, as Google wants, to make the search by barcode easier.

Table 1. Example of a barcode Barcode: RMSSTA$$000000497$$$F

Barcode: RMSSTA$$000000497$$$F	
RMS: polo code	ST: Earth Science Library code
A$$000000497$$$: inventory A 000000497	F: check digit

We decided to create a univocal ID with significant elements (not random univocal sequences) just to have an ID with an own link to the material object.

3.3 Such as One Collection

This process give us a global vision of all the books selected by libraries, so we send without mistakes just one of the items of a work and we don't digitize the same edition two or three times, for example, in two or three different libraries.

From the beginning, we decided not to not duplicate records, because, when we have more copies of a specific edition kept in different libraries, first we have to analyze the conditions of every single book to choose a suitable copy for digitization.

4 Work-Flow

4.1 Selection of the Books for Date and Realization of Lists for Next Steps

Often the bibliographic descriptions in our catalogue don't follow the standard punctuation of cataloguing rules, so the files exported in marc are not necessarily suitable

to select books for year of publication. The solution was to search for year of publication both in marc subfields and in ISBD description field (using a temporary database) with a specific procedure that recognizes frequent mistakes in cataloguing. This way we have more results than just searching in the marc field (10% more).

4.2 Declaration of Suitability According to the Project

We put all the data in a Google spreadsheet, one per library. The files are shared on a website with restricted access that contains all the information about the workflow. The librarian inserts in this file the following information:

- if the book is available
- if its material condition is suitable for digitization process
- book value, calculated by a Sapienza Library community algorithm

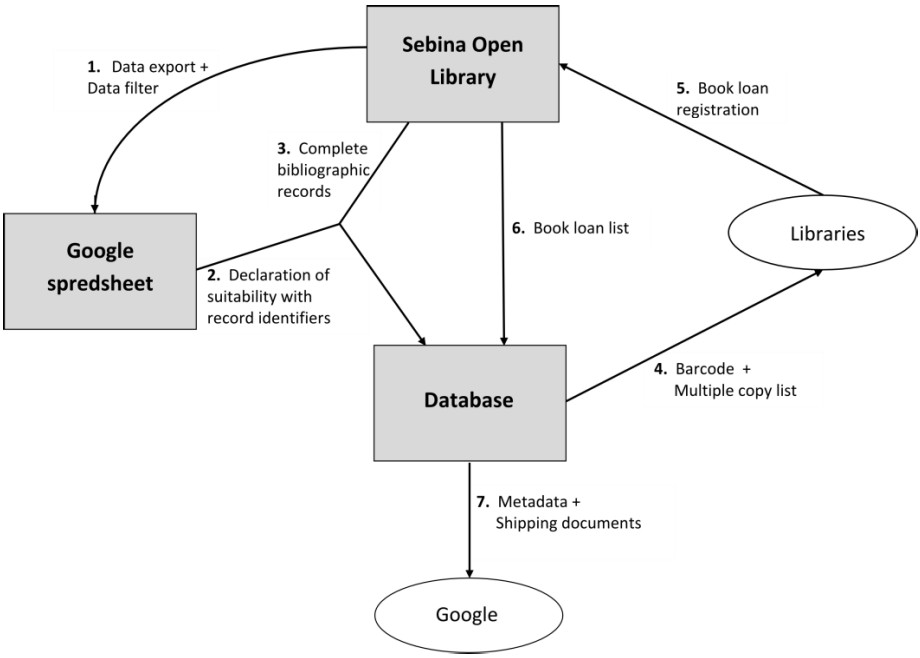


Fig. 1. Diagram of the work flow

4.3 Record Duplication and Barcode Production

All data produced by librarians are imported in a database (MySql) for the next steps. The process that populates the database, developed in Php by Cineca, uses "Google spreadsheet API". Only after the comparison of suitable book data, given by librarians, a unique list can be created choosing just one volume per work (usually the first one inserted in the server, in chronological order); a barcode is linked to the selected

book and all other copies are automatically discarded. The barcode generation phase is therefore also used as a check of multiple copies. The report list management also makes it possible to check if the books have already been digitized in past.

4.4 Books Loan Registration and Sending

When library staff put the books on the cart for shipment, they also register the loan in the catalogue. So, for every cart sent, there is a user ID with a standard code, linked to the library. The temporal sequences of book loans (timestamp information) are used to produce the correspondent lists of volumes ID, separating correctly "ready to go" carts from "come back" carts, for the ten different libraries across la Sapienza and in the city of Rome too. So, the list of volumes registered in Sebina is imported in the database; it is useful to take note of books sent every time, to prepare the necessary documents for each cart and to have a complete check list of all sent books.

5 Results and Developments

For students and professors, especially of this University, the Google Books project means a great improvement of services, both for the access and the quality of checked and enriched data.

The project that we are developing also includes an extension of document typologies for digitization and a better integration with the cataloguing database.

Until now, only monographic volumes were scanned, but we are going to digitize serials too. Serials have some peculiarities: for example, every single issue of a magazine is linked with a bibliographic description and it means a surplus of work is needed to make the digitization management possible and to make the identification of each digital object easy.

The involved libraries are increasing and the direction of next phase is getting clear; so it's necessary, as soon as possible, to align the cataloguing database with all the spreadsheets (now hand-made), maybe by web-service, just as it happened with the alignment of all student databases.

Multimedia Digital Libraries Handling: The Organic MMIR Perspective

Roberto Raieli

Roma Tre University Arts Library, Rome, Italy
roberto.raieli@uniroma3.it

Abstract. This paper focuses on new retrieval methods and tools applicable to the management of multimedia documents in Digital Libraries (DL). These matters merge in the organic methodology of MultiMedia Information Retrieval (MMIR). A paper's goal is to demonstrate the operating limitations of a generic Information Retrieval (IR) system, restricted only to textual language. MMIR offers a better alternative, whereby every kind of digital document can be analyzed and retrieved with the elements of language appropriate to its own nature, directly handling the concrete document content. The integration of this content-based conception of information processing with the traditional semantic conception, can offer the advantages of both systems in accessing of information and documents managed in actual multimedia digital libraries.

Keywords: multimedia information retrieval, content-based information retrieval, multimedia documents, digital libraries, image and video processing, audio processing, indexing, semantic gap.

1 Introduction on Today's Context of MMIR Development

1.1 The Contemporary Panorama of Information

In the Library Science community, new methods and tools of processing and searching for the management of new multimedia documents in actual Digital Libraries (DL) are the emerging issue. DL databases do not store only mainly textual documents, but also documents such visual, audio, audiovisual or *multimedia* in the full sense. This problem is directly linked to issues of disseminating and accessing of documents and information, core objectives of the DL activity, and it emphasizes the need for new multimedia modalities for the treatment of every kind of digital documents, in databases and in the Web too.

In this panorama, a contradiction often arises. It is related to the *terminological* logic by which information systems and services continue to be organized, despite of the radical changes through which documents evolved into multimedia or hypermedia objects. If searching a written document is not possible by visual or sound language means, then retrieving documents consisting of sounds or images using descriptive texts should not be considered as an effective method.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 171–186, 2014.

In today's information and knowledge society the various limitations of operating within the logic and the terms of the general structuring of Information Retrieval (IR) should appear evident. In the traditional practice of IR each kind of document searching is brought about through *textual* language. Now, it is necessary to define the features of a MultiMedia Information Retrieval (MMIR) system, where every kind of digital document can be processed and searched through the elements of language, or *meta-language*, appropriate to its own nature.

Experimentation and use of MMIR technologies are already well developed within computer engineering, artificial intelligence, computer vision, or audio processing fields, while the interest in the methodological and operational revolution of MMIR, and the reflection on its conceptual development, have yet to be introduced among librarians, archivists and information managers. The contexts of the Library and Information Science (LIS) still have the opportunity to welcome the discussion, addressing the development of MMIR systems for DL according to needs of Library Science order, at a time when MMIR databases and interfaces are in the testing phase.

This new vision is really suitable for multimedia digital libraries handling. Four methods within the general and *organic* methodology of the MMIR can be distinguished: a method of Text Retrieval (TR), based on textual information for the processing and searching of textual documents; a method of Visual Retrieval (VR), based on visual data for the processing and searching of visual documents; a method of Video Retrieval (VDR), based on audiovisual data for the processing and searching of videos; and a method of Audio Retrieval (AR), based on sound data for the processing and the search of audio documents.

The different matters developed around traditional systems and services of IR and DL management change entirely when the MMIR point of view is considered. In databases where the content of the documents is substantially text it is appropriate using as access keys terms and strings extracted *from the inside* of that content. However, in databases of images or sounds it appears over-simplified and inaccurate to allow access, *from the outside*, through a textual description of contents that are often indescribable by terms.

Within the MMIR logic analysis and search methods are defined as *content-based*. They are structured on a methodology defined as Content Based Information Retrieval (CBIR),[1] which provides keys of storage and retrieval of the same nature as the *concrete* content of the objects to which they are applied. These keys are based on a language appropriate to every document typology, able to point with congruence to the concrete content, as well as to the aspects of meaning of a certain document.

1.2 The Current Policy of Classification and Indexing

From the MMIR perspective, Information Retrieval is defined as a *term-based* system of indexing and searching. This definition given to the traditional IR system emerges from the new conceptions of CBIR, and it addresses the problem that in IR the use of

[1] In several interpretations CBIR is "Content Based Image Retrieval".

textual language and the methodology of terminological treatment always appear as the natural and only one way to consider documents and information.

Within the traditional organization of libraries, databases and DL, a number of attempts have been made to adapt IR systems to the new demands of users and to the needs of multimedia documents, but these attempts often have resulted in highly complex and difficult solutions. The weakness in common among these experimentations is the incapability of renewing the fundamental principles of the system.[2] What is needed, instead, is a general revolution of perspective, replacing the principle of term-based document processing with the content-based principle, which is adequate to appraising dynamic multimedia content as well as textual content. In fact, the main criterion of the *contentual* analysis of documents is to constitute directly the means of processing, searching and accessing on the basis of the real content of each document: text, figure, sound, or a whole richly combined.

In the specific field of visual arts, innovative thesauri have been established for indexing every kind of image related to various forms of art. One of the more important classification indexes surely is the *Art and Architecture Thesaurus* (AAT) [3]. A second salient art classification system, a partial alternative to AAT, is *ICONCLASS* (ICONographic CLASsification System), applied to an "iconographic" classification of the objects [4]. However, also these attempts can be criticized for trying to resolve the problem within the traditional term-based system.

2 Multimedia Handling of Digital Documents

2.1 The New Way to Documents Searching

In the last twenty years, the growing importance of multimedia documents, the new tools offered by digital technologies, and the creation of multimedia databases and DL of high complexity, have led to investigate the possibility of *multimedia analysis* and *indexing* based on the real nature of multimedia queries, which must address search techniques to operate within the new multimedia digital libraries.[3]

The debut of CBIR, in the late 1980s, was founded on image processing and on computer vision studies [7-8]. Then, in the late 1990s, the attention to video documents progressed, managing visual documents involving movements, speaking and sounds, and pressing the studies toward a more complex MMIR [9]. Therefore, at the beginning of the 21st century, investigating problems related to user interaction and system response has been possible [10], as well as the improvement of processing algorithms [11]. Finally, today, more specific problems can be studied, related to the

[2] Some advanced proposals in the methodology of IR were presented by Nancy Williamson and Clare Beghtol [1]. Another important theoretical reflection on IR is the "pragmatist" issue discussed by Hjørland and colleagues [2].

[3] Elaine Svenonius was among the first researchers to comprehend the problematic new area of indexing languages [5]. William Grosky was among the first to draw some general conclusions for a coherent and effective management of new multimedia databases and DL [6].

semantic understanding of the query, to assure a system able to *understand* user's request through both contentual and conceptual specifications [12].[4]

During this development, the increasing use of IR in commercial and scientific circles has also stimulated specific interest in the field of the Computer Science that, unlike Library Science and Documentation, has faced various problems from the perspective of the processing and evaluation techniques for the raw constitutive data of document contents. From a computer-centred perspective the way consists: in the construction of new and specific indexes of multimedia data, in developing high-level analysis and query systems with many options, in developing data analysis algorithms able to calculate a huge number of variables, in the setting up of results evaluation and ranking systems that improve response quality also interacting with user specifications, and, finally, in the development of analysis and search paradigms able to relate the *automatic* objective representations of the computer with the *intellectual* sophisticated analysis by the human.

Anyway, the state of the art of MMIR systems still shows a series of open problems, with several consequences [13]. The main problem is imposing the content-based method for multimedia information processing having such advantages that it will naturally replace the traditional IR system. To establish a utility-centred research focus is critical, bridging the so called "utility gap", or the distance between users' expectations and real systems usefulness [14]. Specific methods and protocols of evaluation and benchmarking for MMIR systems are necessary, allowing appraisal of advantages and ineffectiveness, of user's satisfaction related to procedures and results, and of improvement possibilities.[5] So, one of the great challenges for the future is the need to move from the academic and experimental state of MMIR to a practical and commercial phase, based on cooperation between research and industry.

Beyond this, since the effectiveness of the information process is largely influenced by the *interaction* of the operator with the system, a lot needs to change relating to the user. The whole system of approach to multimedia databases and DL must be re-established on the basis of the natural and increasing demands to define the query by operations in continuous interaction between human and computer [15]. Various researchers are occupied with analysis of surveys taken in documentation centres, libraries or archives, focused on the verification of the usefulness of MMIR interactive methods, and of the active learning of the system arising from user's relevance feedback [10, 16].

Among studies about MMIR effectiveness for users, the most successful line is the English one, in which the work of Peter Enser and Christine Sandom is predominant [17]. This has brought CBIR researchers to stigmatize as a "semantic gap" the semantic ineffectiveness of the search systems based only on automatic content processing. Such a void is identified in the distance between the *high-level* conceptual representation of an object, appropriate to human knowledge, and the *low level* formal

[4] See also the web site of the Semantic Media Project, announced in 2012: `http://semanticmedia.org.uk`

[5] See the TRECVID activities on the web site of TREC Video Retrieval Evaluation: `http://www-nlpir.nist.gov/projects/trecvid`

denotation, appropriate to the computer. Therefore, the semantic approach cannot be neglected by a content-based system, and the necessity that a complete MMIR system allows every search through all the means that the user desires – semantic and contentual – is once more confirmed.[6]

Finally, very relevant for the stabilizing and the growing significance of MMIR studies, is the foundation in 2012 of the *International Journal of Multimedia Information Retrieval*, aiming to present achievements both in semantic and in contentual treatment of multimedia.[7]

2.2 Principles of MMIR and Content-Based Retrieval

Many of the strings that users create to query a multimedia digital library or a multimedia database, or also the Web, are aimed for a search that goes beyond the information or subjects definable with precise term constructions, and points to qualities appropriate to the content taken *in itself*. Simple queries, not subsequently refined relating to time and space, to actions, to the expressive forms, can be satisfied in the ambit of term-based systems. However, more complex query strategies require completion with further operations that by the traditional methods and tools do not always bring the result that the user expects, or they are simply impossible.

Therefore a system of MMIR is more helpful, since the formulation of the query does not have to be forced within the limits of the textual language, but it can be sent as it is naturally produced, directly in visual, sound, audiovisual, and textual means. This is a really new model, where the query can be expressed to the system as it arises, and as it arises it can be appropriated and answered by the computer, according to a content-based processing logic: through colours, forms, structures, sounds, movements etc. – and words, when they are the content.

This will be possible only if documents are analyzed and indexed not only according to the terminologically reportable or translatable data – semantically – but also by structuring a sort of index directly constituted by the concrete and formal data – contentually. However, in this context the concept of *indexing* must be intended in wider sense: a content-based index will be *made* of the data with which the computer operates for reproducing images, sounds, or words contained in the documents.

The sense of the problem can be schematized with a simple example of MMIR and, especially, of Visual Retrieval. A search system that imposes to set terminological strings is not useful for someone who desires to retrieve some images having a certain combination of forms and colours. Any combination of phrases will fail the retrieval goal if only the name of the author, or the title of the work, are in the set of the indexing terms. Indexing or classification data refer to another system setting, of an intellectual and specialist kind, and they seem to be abstract data relating to the image, useful only when they are known before the search (e.g. fig. 1).[8]

---

[6] However, some researchers have expressed different opinions. For an overview of studies on users' needs, see also pp. 231-234 of Frederick Lancaster's book [9].

[7] See the first papers of: International Journal of Multimedia Information Retrieval. Springer, London (2012-).

[8] The painting is by Leonardo D'Amico (*L'ultima neve*, 1998, oil on canvas, 30x50 cm).

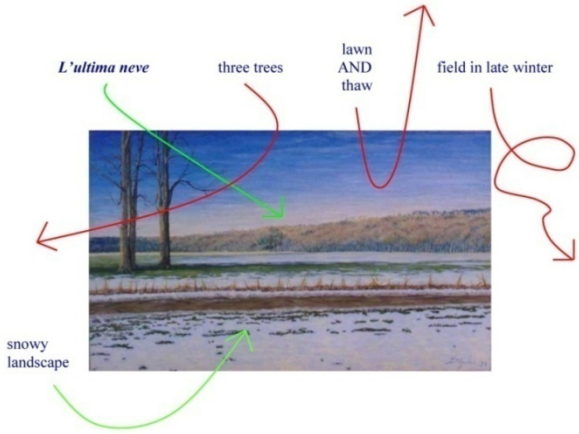


Fig. 1. Example of textual-visual search

If a VR system can be searched by proposing the combination of textures or shapes and colours that the user imagines, or he vaguely remembers, it is possible to go directly to the contentual *core* of the document concretized by the image. In this way, a visual document can be retrieved together with similar documents, and with information, textual and conceptual, connected in several ways to it (e.g. fig. 2).[9]

Five different levels of VR processing for visual documents can generally be carried out. The *semantic* mode is the most traditional method, and it consists in

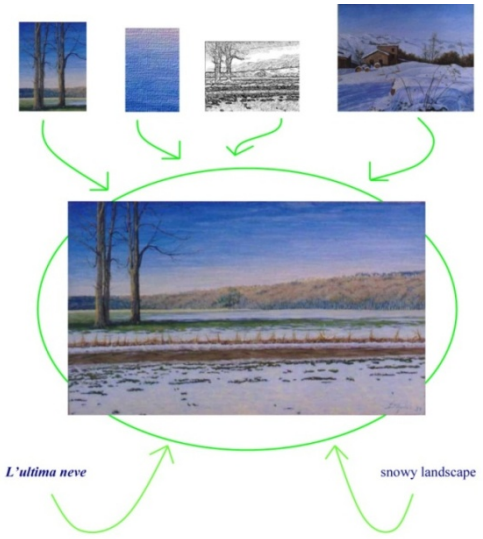


Fig. 2. Example of visual-visual search

[9] Top right painting by Leonardo D'Amico (*Silenzi invernali*, 1998, oil on canvas, 40x50 cm).

defining text labels, which describe characteristics, classes, meanings, titles or names, attributed to an image. The *shape* retrieval mode relies on the computer's ability to compare extracted forms of an archived figure and those extracted from a query model. *Texture* processing is based on breaking up stored images into sections, then the system estimates the similarity of the structural composition with a model figure. *Colour* processing consists in representing images using colour or grey scale properties. Finally, the *parametric* mode is based on determining parameters of shape, texture and colour of images, through figure templates or by filling in a grid [12].

Video Retrieval documents treatment has some in common with VR, but handling audiovisuals needs to give consideration to elements such as time, movements, transformations, editing, camera movement and, often, sound and text data. VDR processing runs by the extraction of *video-abstract* characterized by spatio-temporal factors, supplemented by information on textual data relating to the written and the spoken in the video.

The first VDR treatment operation is usually the rearticulation and *segmentation* of the video stream into four levels of increasing complexity: the frame, which is almost always a still image; the sequence, which is an early articulation of frames in spatio-temporal development, and it may have sound; the scene, that has a high level of complexity, in which sequences are connected to create a sense; the entire film, that is a unique product of all scenes, giving sense and meaning to the whole. After video segmentation there are analysis and extraction of the so-called video-abstract, or video-summary, that is a base for query and retrieval processing as it is of less complexity than the entire video. Queries can be set by key-frames, allowing users to launch a search in form of visual query. Information on movements and sounds, as texts, can then be added for the completion of the video-query [18].

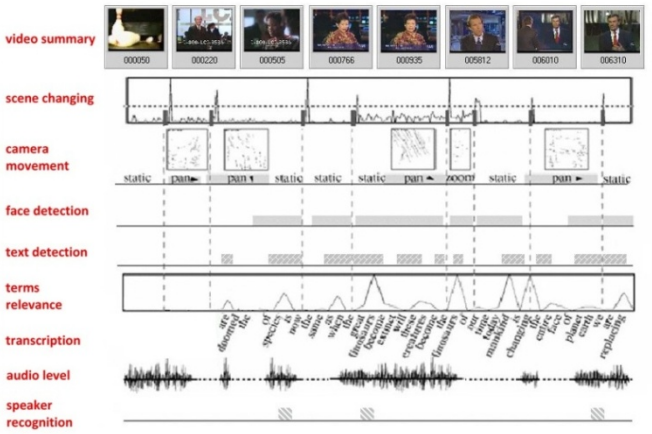


Fig. 3. Model of a video elements analysis

Audio Retrieval method differs since an audio data stream is connoted by *tempo-related* properties, and properties relating to frequency and sound characteristics such as tone, pitch, timbre, melody and harmony. The same can be said for audio

documents treatment, as AR techniques have some in common with the whole MMIR, but specialising under specific sonorous respects. This even means working directly with contentual elements and concrete objects, as *ineffable* as sounds may seem, without excessive mediation based on terminological methods.

Emerging issues in AR are robust notation, alignment of different versions, comparing variations. Emphasis must be given to audio-thumbnailing, or making audio-abstracts, and audio browsing. They are connected together, in as much as drawing a complete effective sound synthesis allows to browse and evaluate an audio document in the analysis and search. Some running modalities of AR are: speaker identification, based on the ability to recognize human voices regardless of the words spoken; the typical *similarity query*, which search a database using sample tracks; query-by-humming, that allows search by similitude to an audio model or strummed or hummed by user. Main models of AR may be considered Music Information Retrieval, Music Recommender Systems, and Automatic Speech Recognition [19].

3 Bases and Sense of the MMIR Organic Approach

3.1 Objectives and Effectiveness of the Content-Based System

The scope and the goals of the whole content-based system can be specified starting from a schematizing of the development from IR principles to MMIR principles: the Information Retrieval is a system of analysing and searching, through *terms*, of documents of textual kind, which also can be applied to visual, audio and video documents; the MultiMedia Information Retrieval is proposed as a general system of processing and retrieving, through *texts*, *images* and *sounds*, for documents of every kind or full multimedia.

The content-based processing method is proposed as the only one really able to achieve the goal of the MMIR: to retrieve the object that is truly being searched, beyond any abstract mediation of a linguistic and intellectual kind.

Finally, considering the sense of the *organic complex* of the four MMIR specific methodologies – TR, VR, VDR and AR – it is necessary to specify that, to reach a good level of precision in the retrieval of multimedia documents, all the search modalities need to work in constant interaction, inside a single system, according to a univocal principle. A single search interface is required, allowing the composition of a query formula that combining images and texts, sounds and terms, or all these together, is useful for searching very complex documents – whose informative content extends beyond all the levels of *sense* and *meaning*, where the semantic definitions do not have less importance than the contentual characteristics.

In a multimedia DL, by a common MMIR search procedure, the user may start the query with the preliminary selection of a section, and of a part of it, by using appropriate text strings. Using definitions, terms, titles and names can be a suitable and fast method for reducing the massive content of the whole database, as to correct some result *noise*. The user may proceed with browsing, by which queries can be simply sent to the system by selecting some of the retrieved objects, or by assembling or inputting from the outside example models. In this way, by moving among so

many objects that resemble the desired one, or in relation to it, it is possible to send to the computer different example data, containing the characteristic data that must be matched with the objects in the database in order to extract them as query results.

3.2 Mathematics and Sensibility

A critical question of MMIR is what effectiveness the mathematical procedures of content-based systems can have, relating to the practical objectives of users' information searching. The research for computational algorithms and data processing techniques which are not only mathematically *efficient* but also pragmatically *effective* goes toward overcoming of the distance between computer and human, taking into account the information qualities expected by the human operator [20].

The more technical-theoretical problem is the *interpretation* of the multimedia object. This has a considerable value in the search process when the demand of information goes beyond the perceptive characteristics of the object – automatically calculable by the computer – and goes to the level of the semantic interpretation – definable only by the human. So, the content-based query needs to be knowledge-assisted, which means that the user may query the system with the support of *subject* descriptions. The use of subjects created by the human operator can be very useful to indicate both to the user and to the system what the mathematical analyses of an example model cannot directly gather. Therefore, even if the mechanical and absolute efficiency of the mathematical processes may be certain, their utility related to the demands of every end user is not certain.

The automated procedures of MMIR systems avoid superfluous mediation, handling directly the original object characteristics or, more exactly, the data of its digital *version*. Moreover, the data into the search index can directly be produced by the system that will use them, in the more appropriate manner, including operations speed. Possible algorithm's mistakes or approximations are due to known causes, and they are calculable as systematic errors that can be taken into account in managing the final results.

So, in comparison to the individual variables of manual methods and to the hidden interpretation errors, the automated processes are often of greater reliability, at least in certain contexts. Nevertheless, very advanced and expensive hardware and software systems are necessary for an effective processing of larger and richer multimedia documents, and this surely slows the investigation and application of new retrieval technology, while the consequent advantages are unclear.

Primarily, however, content-based and automatic methods are not always appropriate to satisfy the increased demands of researchers and experts, as of common users. The sense of an object represented in a document, in fact, has to be gathered in its *true totality*, in the simultaneous consideration of its sensitive and intellectual qualities. Systems oriented to the concrete content are inadequate to indicate the multiplicity of the intellectual interpretative points of access, and the nonexistent sensibility of the computer cannot fully be produced by algorithmic elaborations of the numerical data representative of the qualities of the documented objects.

If MMIR systems succeed, somehow, in showing validity in the case of a direct and *contentual-objective* approach to the document, they present a greater narrowness in the case of a theoretical and *intellectual-interpretative* approach. Besides this problem, well known as the *semantic gap*, there is also to be considered the parallel problem of the *semiotic gap*, or "sensory gap" [21].

3.3 Integration of Contentual and Semantic Principles and Methods

The solution to the conflict between conceptual and concrete means of accessing to information, or between term-based and content-based systems of processing, can only be a solution of *organic integration* between principles and methodologies.

Such an ultimate achievement has been in development especially by Peter Enser [22]. Comparing the two search methodologies, the author does not use the definition "term-based" anymore, since this has been abundantly criticized, but rather he speaks of a possible "alliance" between "concept-based" and "content-based" paradigms. Enser proposes a technical-practical solution represented by the "hybrid systems" of image retrieval. The search interfaces of such systems allow the "terminological formulation" of the query, "text-matching" of documents based on terms contained in metadata, CBIR techniques to input "concrete search models", and evolved modalities of "translation" of a terminological query in visual query.

However, searching for a true organic principle for the MMIR method cannot be resolved through a simple hybridization of techniques. The limits of the contentual-objective consideration of the document and the discrepancy in comparison to the semantic-interpretative consideration are the explicit manifestations of the problem of the semantic gap. The purpose of MMIR system is to give the support to overcome such voids through the *simplicity* of document processing offered by the computer and the *rich* semantic expectations of the user.

Jonathon Hare, Paul Lewis, Peter Enser and Christine Sandom stress the characteristics of such a gap of representation [23]. The representative levels of a document vary from the lower level, composed by the simple extraction of its "raw data" immediately extracted by the computer, up to the higher level, constituted by the "semantics" that it carries as they are interpreted by the users. When the meaning is considered, in addition to content, this opens a void between the lower and the higher levels in which the documented objects can be positioned.

A satisfying proposal for a solution, given by Enser and colleagues, is "to attack the gap from above", considering the use of *ontologies*. A large set of annotations and labels related to an object is far from representing it in its semantic richness, which seems, instead, to be representable positioning the object inside an ontology. The appeal to ontologies in MMIR systems allows making *explicit* part of the meaning of a document, and this makes possible formulating the query also through the concepts and the relations among concepts. Thus, the multimedia query can be semantically completed, integrating content-based search tools [23-24].

Using ontologies is the way to establish an organic approach for all multimedia document kinds, able to take into account univocally their concrete and conceptual representation. Besides that, however, some other considerations are necessary which

concern one of the fundamental principles of the MMIR: *imagination* and *creativeness* as a *style* of the method to conduct information and documents searching. Accepting the integration of ontologies in MMIR systems, a certain *rigor* seems to be residual in these conceptual tools, which can raise again the problem of the rigidity and abstractness of the typical IR schemes. To avoid such risk, a further hypothesis can be made about combining ontologies with *folksonomies*, systems of free collaborative categorization of contents on the basis of labels directly assigned by end users, or "social tagging". These systems, proposing their function close to the controlled semantics, can enrich them of more flexibility in metadata and tags definition. In this direction goes a discussion started by the same founders of the Semantic Web and Web 2.0 [25-26].

Everything is abreast of the principles of the MMIR, where the possibility for the user to search freely through models or sketches allows the system to *learn* at the time new information on the documents, that will be stored together with information already defined, integrating and widening its *interpretative* abilities. The integration between the semantic tools of ontologies and folksonomies, contemporarily integrated to the content-based tools of CBIR, can bring to the reconciliation of the opposition between the principles of the semantic-interpretative and of the contentual-objective information handling, in the general organicity of all the organs of the MMIR.[10]

4 Definition of the MMIR Methodology and Conclusions

4.1 MMIR Paradigms Currently Being Studied

Paradigms and protocols tested so far in design and implementation of MMIR methodology are all quite similar. In general, in the system process two major interrelated parts can be distinguished: operations relating to the documents analysis and the creation of databases and indexes, and operations concerning the processes of search and retrieval of documents and information [10-11, 20].

As regards the content-based treatment and analysis of documents and the consequent creation of databases and representative indexes, some steps are required, but they are not necessarily sequential and are often repeated, updated or integrated with each other. These steps can be summarized in table 1.

Before the system can effectively apply to the content-based processing, the pre-processing analysis of the multimedia files is crucial. Multimedia data must be treated according to multiform strategies, capable of detecting also information related to rich structures or continuous changes of objects. Such a characterization may be conducted automatically, saving time and costs, and then almost always must be integrated with human intervention.

If the analysis of the semantics needs to be broadened, some intellectual interpretation is required. However, human intervention can be deferred until some syntactic features of an object can be used. For example, in an advanced video

[10] A broader discussion of the gap and possible solutions is in an author's previous paper [27], focused on discussing the problem of the semantic gap in new multimedia search methods.

Table 1. Analysis, storage and indexing

Analysis. Before introduction into the system, documents are processed with its analysis tools, automatic or semi-automatic, to identify the elements of their content.
Datafiling. The constitutive elements of each document are elaborated for creating the general data file, representative of the whole object, stored in the system database.
Characterization. The characteristic data of the main aspects of a document are drawn out of its content, from the data of the constitutive elements. The characteristic data are, then, inserted in the metadata connected to the general datafile of the document.
Indexing. The index is created and updated constituted by the characteristics and general data of the documents. Every value is represented once only but, for each value, a referring link is created to every datafile and metadata that contains it.
Description. Documents characteristics are valued, in a manual, semi-automatic or automatic way, and then described through numerical or textual strings. These strings are inserted in the metadata of each document and represented in the index.
Interpretation. The contents of the documents are semantically interpreted by the human operator, assisted by the system or not, to identify their conceptual aspects. The various semantic indexing terms are inserted in the metadata and in the index.
Query analysis. The analysis operations of the document can be entirely executed in automatic way, also during the query, to allow very free and interactive searches.

analysis, movement can be considered to add meaning to the material characteristics of a figure. Then, since the analysis and indexing processes are never definitive, even the learning tools applied in a query are useful to the system to classify documents, automatically learning from semantic information spontaneously sent by the user.

Content-based characterization and description are stratified in levels. At the lowest level of extraction are mere pixels, representative of shapes, colours etc. In an intermediate stage, complete objects and their content characteristics are extracted. At the highest level, abstract concepts can be derived from the document, forming the human interpretation. Syntactic representations, such as histograms and structures, are indicative of the data extracted in the lower levels. Semantic descriptions, such as labels, are meaningful to the abstractions in the higher levels.

The indexing process is distinguished in two operating levels. The syntactic indexing level allows searches based on templates or via a sample, which forces users to the data extracted from the lower stages of analysis, as details of texture, shape, size and so on. The semantic indexing level fosters searches also via conceptual elements, but according to current technology this indexing not necessarily needs be produced by human intervention, as automatic tools may calculate probability and recurrence of elements able to give a first objective interpretation of documents.

Even the several steps of structuring search operations and documents retrieval do not necessarily have a preparatory order, and can turn around themselves during the processes. These operational phases can be summarized in table 2.

Table 2. Search and retrieval

Preliminary search. The first approach to a system usually consists in a terminological interrogation, through texts or through selection by menu and lists, with the purpose of selecting a part of the documents in the whole database.
Model composition. A tool for the creation of query models allows the creation of an example of the desired object with which to start searching the system.
Model proposal. It is possible to propose external models for the search that allows interrogation by models introduced from outside the database.
Search. The core phase of the search consists in the use of the identified documents, or of the proposed models, or of various single elements, as data for the comparison operations with the data of the index related to the objects of the database.
Matching. The system detects the match between the search data and those of the documents in the database when their similarity is included in the planned evaluation parameters. Then the system achieves the automatic capture of identified documents.
Ranking. When a number of documents are captured, the system shows them in order from the more to the less correspondent to the different required characteristics, and this allows the user to browse and to value them.
Deepening. After the evaluation of the first results, the search can be deepened using further extracted documents, changes in the objects characteristics, selection of their parts, or the association of various contentual and semantic elements.
Interaction. All the operations are often in a phase of interaction and of learning. The interaction with the operator allows the system to understand the user's search criterions and to address the query, so the computer learns from the human interpretation given to the different steps of the process.

The main parameters of the matching techniques are the level of characteristics extraction from the documents and the structural measurements. The common search and retrieval strategy found in different systems is almost always based on the low-level characteristics of multimedia objects, without any ability to implement automatic interpretations of them, assigning to users the task of defining the relevance/non-relevance of certain document characteristics.

A real interaction with the user, however, allows the system to understand humans' search criterions and to address the query. Emerging learning methods foresee a system able to learn by users' spontaneous instructions produced during the search. Automatic data acquisition can be used by the system to build models of categories or domains that will be referenced in the automatic evaluation of an object as if the system had learned an *idea* of it. Difficult goal is to put in place a search and retrieval framework including semantic indexes supervised or created by the human.

Developing robust relevance ranking algorithms is also crucial. So, the system may show retrieved documents in a reliable order from the more to the less correspondent to the query goal. This allows the user to browse and to value search models further, also using terminological means not only for preliminary search, thus deepening again the query between system results and semantic evaluations.

4.2 Concluding

The context of the MultiMedia Information Retrieval is composed of the traditional systems of analysis, search and retrieval of textual, visual, sound and audiovisual documents, of today's systems applied to the management of new multimedia documents, and of the different theoretical and technical attempts to establish a more advanced and effective system for handling, organizing and disseminating the whole of digital documentation in digital libraries and in the Web. This context is varied and dynamic: it involves the work of very different professionals, but it can be interconnected by the goal of realizing a simple and effective organic information tool, which answers the demands of as many users as possible.

Inside the *organic* set of the MMIR another complex is present, composed of its *organs* endowed with specific theoretical, technical and applicative aspects, and appropriate for every kind of multimedia document search: Text Retrieval, Visual Retrieval, Video Retrieval and Audio Retrieval.

The more advanced MMIR systems can be very useful in the support of both theoretical research and creative practice, as well as a tool for professionals and a guide for general users. The user's query can simply and freely be constituted by the input of a model image or sound, with or without conceptual specifications through parameters or texts, and the system can retrieve documents that possess similar characteristics. The user always can interact with a system predisposed for welcoming unpredictable variations of the search way and for *understanding* the human strategy, learning time by time from the researcher's behaviour.

Concerning the organic complex of the MMIR methodologies, in order to reach a good level of precision, the coexistence of all modalities of retrieval is essential, including those based on terms. The terminological query is useful as a preliminary method to select part of a large database and to centre the search basing on data such as information ambits, typologies, classes, titles, or authors. Then, it can constitute a system for cleaning the inevitable *noise* of a content-based interrogation, by specifying a semantic interpretation that the automatic system is not able to detect in the direct analysis of the content characteristics of the document.

All the different procedures operate better in continuous and organic interaction, in a single query interface. Allowing several search strategies, combining words, figures, movements, sounds and concepts, is critical for searching very complex documents, whose content extends through all levels of *sense* and *meaning*.[11]

References

1. Williamson, N., Beghtol, C. (eds.): Knowledge Organization and Classification in International Information Retrieval. Haworth, New York (2003)
2. Hjørland, B., Nissen Pedersen, K.: A Substantive Theory of Classification for Information Retrieval. Journal of Documentation 61(5), 582–597 (2005)

[11] For some web examples of MMIR resources, see: MediaMill, `http://www.science.uva.nl/research/mediamill`; MILOS, `http://milos.isti.cnr.it`; QBIC, `http://wwwqbic.almaden.ibm.com`; QuickLook, `http://projects.ivl.disco.unimib.it/quicklook`; SoundFisher, `http://www.soundfisher.com`

3. Molholt, P., Petersen, T.: The Role of the Art and Architecture Thesaurus in Communicating About Visual Art. Knowledge Organization 20(1), 30–34 (1993)
4. Grund, A.: ICONCLASS: On Subject Analysis of Iconographic Representations of Works of Art. Knowledge Organization 20(1), 20–29 (1993)
5. Svenonius, E.: Access to Nonbook Materials: The Limits of Subject Indexing for Visual and Aural Languages. Journal of the American Society for Information Science 45(8), 600–606 (1994)
6. Grosky, W.I.: Managing Multimedia Information in Database Systems. Communications of the ACM 40(12), 73–80 (1997)
7. Enser, P.G.B.: Pictorial Information Retrieval: Progress in Documentation. Journal of Documentation 51(2), 126–170 (1995)
8. Del Bimbo, A.: Visual Information Retrieval. Kaufmann, San Francisco (1999)
9. Lancaster, F.W.: Indexing and Abstracting in Theory and Practice. University of Illinois-Graduate School of LIS, Urbana Champaign (2003)
10. Sagarmay, D. (ed.): Multimedia Systems and Content-Based Image Retrieval. Idea Group, Hershey (2004)
11. Adami, N., Cavallaro, A., Leonardi, R., Migliorati, P. (eds.): Analysis, Retrieval and Delivery of Multimedia Content. LNEE. Springer, Heidelberg (2012)
12. Enser, P.G.B.: Visual Image Retrieval. In: Cronin, B. (ed.) Annual Review of Information Science and Technology, pp. 3–42. AAIST, New York (2008)
13. Mittal, A.: An Overview of Multimedia Content-Based Retrieval Strategies. Informatica 30(3), 347–356 (2006)
14. Hanjalic, A.: New Grand Challenge for Multimedia Information Retrieval: Bridging the Utility Gap. International Journal of Multimedia Information Retrieval 1(3), 139–152 (2012)
15. Linckels, S., Meinel, C.: E-librarian Service: User-Friendly Semantic Search in Digital Libraries. Springer, Berlin (2011)
16. Thomee, B., Lew, M.S.: Interactive Search in Image Retrieval: a Survey. International Journal of Multimedia Information Retrieval 1(2), 71–86 (2012)
17. Enser, P.G.B., Sandom, C.J.: Towards a Comprehensive Survey of the Semantic Gap in Visual Image Retrieval. In: Bakker, E.M., Lew, M.S., Huang, T.S., Sebe, N., Zhou, X.S. (eds.) CIVR 2003. LNCS, vol. 2728, pp. 291–299. Springer, Heidelberg (2003)
18. Chaisorn, L., Manders, C., Rahardja, S.: Video Retrieval: Evolution of Video Segmentation, Indexing and Search. In: 2nd IEEE International Conference on Computer Science and Information Technology, pp. 16–20. IEEE, New York (2009)
19. Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.: Content-based Music Information Retrieval: Current Directions and Future Challenges. Proceedings of the IEEE 96(4), 668–696 (2008)
20. Maybury, M.T.: Multimedia Information Extraction. Wiley-IEEE, New York (2012)
21. Smeulders, A.W.M., Worring, M., Santini, S., Jain, R., Gupta, A.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analisys and Machines Intelligence 22(12), 1349–1380 (2000)
22. Enser, P.G.B.: Visual Image Retrieval: Seeking the Alliance of Concept-Based and Content-Based Paradigms. Journal of Information Science 26(4), 199–210 (2000)
23. Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval. In: Chang, E.Y., Hanjalic, A., Sebe, N. (eds.) Proceedings of Multimedia Content Analysis, Management and Retrieval 2006, pp. 75–86. SPIE, San Jose (2006)

24. Mallik, A., Chaudhury, S.: Acquisition of Multimedia Ontology: An Application in Preservation of Cultural Heritage. International Journal of Multimedia Information Retrieval 1(4), 249–262 (2012)
25. Shadbolt, N., Hall, W., Berners-Lee, T.: The Semantic Web Revised. IEEE Intelligent Systems 21(3), 96–101 (2006)
26. Guy, M., Tonkin, E.: Folksonomies: Tidying Up Tags? D-Lib Magazine 12(1) (2006), http://www.dlib.org/dlib/january06/guy/01guy.html
27. Raieli, R.: The Semantic Hole: Enthusiasm and Caution Around MultiMedia Information Retrieval. Knowledge Organization 39(1), 13–22 (2012)

Closing the Gap: Interdisciplinary Perspectives on Research and Education for Digital Libraries

Anna Maria Tammaro[1], Vittore Casarosa[2], and Donatella Castelli[2]

[1] University of Parma
[2] ISTI-CNR, Pisa

Abstract. Two major themes continue to be a subject of discussion when dealing with digital libraries: how should the education programs in LIS (Library and Information Science) schools be changed or updated in order to provide the needed knowledge (skills ?) for librarians in the digital age and, closely related, how could the three major memory institutions (libraries, archives and museums) define common educational curricula for professionals in the three domains, now that the digital age is blurring the boundaries between the three profession. In this paper we will present some considerations about the first topic, in order to share the experience gained through the organization and the participation in five events, having as theme the educational needs of the new librarians and the possible synergies of research and education in the field of digital libraries. It is hoped that it can serve as a further stimulus for discussions and for the definition of possible common actions in the digital libraries community.

Keywords: digital libraries, education, research in digital libraries, information professionals.

1 Introduction

Computer Science and Library and Information Science communities practice and do research differently and as a result their outcomes such as curricula, projects results, digital products and publications are different. But digital libraries and the things they bring with it, such as curation of digital collection, interoperability, metadata, which are prompting a move from a "Library model" to a "Digital Library model" are pushing to close the gap between the two communities. This paper examines some of the theoretical differences between the two communities as well as the experiences of sharing expertise and how the Digital Library model is contributing to this overlap and how education and research on digital libraries are evolving to support the new synergies.

In a simple view, the notion of Digital Libraries involves some combination of multimedia content and computer programs. It has unique advantages such as very low marginal costs for creation, storage, management and speed access and distribution but also involves the disadvantage of increased legal obstacles for access to information and the weakness of economic sustainability. Research in "Digital Libraries" has been going on now for over 15 years (even though there is not yet a general

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 187–197, 2014.

agreement on the exact meaning of the term), leading the way to research also in fields of memory institutions. As a result the Digital Library Universe is a very complex one, encompassing a number of different technologies, disciplines and application fields. In addition to that, research in Digital Libraries can be tackled from many different perspectives and angles. Digital Libraries are, for example, information systems and their technology can be researched as such; but they are also organizations and they can be researched also in that respect; they are arenas for the study of information seeking behaviour and for social processes such as learning and knowledge sharing, which can be another dimension of research; they are collections of content that need curation (collection, description, preservation, retrieval, etc); they are social institutions with a social mandate, and as such they are affected by social, demographic and legal issues. Interdisciplinary perspectives cover a wide range of digital libraries management issues and research findings offer insight into educational curriculum and real world practice.

From this multifaceted perspective it appears that Digital Libraries continue to be a new topic in existing research fields, and education has to take into account this interdisciplinary and multidisciplinary aspect. Experts from the two communities should offer their views in the operational, managerial and strategic challenges that face digital libraries managers and researchers now and in the next decades.

2 Literature Review

Taken in isolation from each other, Library and Information Science and Information Technology approaches have a number of constraints. Coleman [7] noted that for too long, LIS schools have responded to the impact of IT in the workplace by simply adding to the existing LIS curricula courses such as Systems Analysis and Design, Database Fundamentals, Human Computer Interaction, and so on. The IEEE Technical Committee on Digital Libraries (TCDL) promotes research in the theory and application of digital library technologies. Issues of interest include: Searching and browsing; Indexing for multimedia objects; Authoring, Scripting and capturing systems; Resource discovery; User interface; Collaborative research; Information representation; Intelligent agents; Workflow; Telecommunication and networking; Interoperability; Scalability; Content storage and distribution; Protection of intellectual property and user privacy; and Accounting, billing and payment systems. The Computing Curricula [8, 10] outline Digital Libraries as an elective area with topics such as digitization, storage and interchange, digital objects, composites and packages, metadata, cataloguing, author submission, etc.

Another approach has been to merge; often the merger is with larger departments such as Communications and Education and less often with IT-intensive ones such as Computer Science [12]. Coleman [7] concludes that anecdotal evidence suggests that both approaches leave novice LIS graduates with overwhelming feelings of information overload, the impression that the library profession is in chaos, and a sense that there is no real core LIS disciplinary knowledge beyond the service ethic, descriptive and procedural knowledge of information resources and their use.

Tennant, a professional librarian, discusses [21] the shortage of digital librarians and explains why public service LIS professionals must become ”tech-savvy”. How can you offer good public service, he asks, if you don't know the "universe of possibilities"? A digital librarian should distinguish ASP from PHP (two different ways of creating dynamic web pages), and be able to understand and evaluate a variety of information technologies for their potential use. Are librarians still needed? Google has become a nearly omnipresent tool of the Internet, with its potential only now beginning to be realized. Users are more and more starting their research from Google page and librarians can become an outdated species. Miller and Pellen [14] comprehensively explore the path libraries need to travel to benefit from the search tool, rather than being overwhelmed and destroyed by it.

Over the past years, digital content has been generated faster than our ability to manage, preserve and disseminate it. Some of the current efforts in research have been focused on improving our capacity for better managing repositories, for preservation and for building infrastructures for searching, accessing and re-using networked digital resources. Bruce [2] affirms that the intellectual and technical issues associated with the development, management and exploitation of digital libraries are far from trivial and we are still a long way to consider it solved. What is needed is a coordinated approach to digital library research combining expertise of LIS and Computer Science with applications such as e-learning, e-government, e-science and digital humanities. This will make it possible to make significant progress towards semantics based multimedia knowledge networks.

3 Methodology

While interdisciplinary convergence is needed, it will not suffice in overcoming all the constraints. We want to share here our experiences of the participatory nature in Digital Library curriculum design and discuss how, as a team with different backgrounds (Humanities and Computer Science, Education and Research), we developed a common understanding using a “workshop model” which has been run and iteratively refined at five major international conferences, involving over 200 participants. The cooperation started with a workshop held in 2005 in Parma with the title “Information Technologies profiles and curricula for libraries”, jointly organized by the DELOS Network of Excellence, the European Library Automation Group (ELAG) and the University of Parma International Master in Information Studies [11]. The second event was in 2008, with a panel organized at the ECDL Conference in Aarhus, having as subject: “The Web versus Digital Libraries: time to revisit this once hot topic” [6]. In November 2010 the DL.org project joined forces with the International Master “Digital Libraries Learning” (DILL, a Master Programme funded by the EC's Erasmus Mundus program), organising a seminar in Parma with the title “Education and Research in Digital Libraries” [9]. In 2011 a Workshop with the title “Linking Research and Education in Digital Libraries” was held at the TPDL Conference in Berlin, as a continuation of the previous one [5]. Finally, a last workshop with the title

Fig. 1. A tag cloud for Research and Education in Digital Libraries

"Can Research help Education in Digital Libraries?" was organized, in connection with the LIDA 2012 Conference in Zadar [1].

In section 4 we will briefly present the main aim and outcomes of the first two events, where the focus was mostly on understanding the educational needs of future librarians, to cope with the increasing need of skills in Information Technology. In section 5 we will briefly describe the other three events, where the focus was mostly in trying to understand the possible sinergies of a closer cooperation between Research and Education in the field of Digital Libraries.

4 Professional Profiles for Digital Library

Since the impact of IT on all disciplines will only continue to increase, also the interest for interdisciplinary curricula development is increasing. A new approach should view curriculum development intellectually at the unit level (what topics and learning objectives/competencies are common across related disciplines) and how best to facilitate this development for professional graduates. At the very least, such approaches could use research findings about interdisciplinary learning to improve the problem solving and competencies of graduates.

The Workshop "Information Technologies profiles and curricula for libraries", held in Parma in 2005, wanted to identify and define the IT profiles and skills needed in libraries and information centres, and to propose a set of educational actions that could result in making those skills available in the short medium term. A first set of presentations was on the theme "Relationships and cooperation between IT education and LIS schools", which provided an overview of the current position in educational issues with respect to information technologies for libraries and recent trends in research on digital libraries. A second set of presentations was introducing and

discussing the general theme of “Competencies and Profiles”, devoted to the contributors' own experience and case studies of skills and competencies. Both sets of presentations were supplemented by a process of feedback through a series of parallel breakout sessions and workgroup discussions, which were then reported back and discussed in plenary sessions on the second day.

During the workshop two “new” professional profiles needed in a digital library were discussed. The first profile was that of a “digital librarian”, with a deep knowledge of the (digital) content of the library, and enough knowledge of IT tools to allow him/her to” curate” (the term was not yet trendy at that time) the collections of the library. The second profile was that of a “system librarian”, with a good knowledge of Information Technologies and Architecture, and enough knowledge of library services and management to allow him/her to formulate the requirements for a Digital Library Management System and to use and manage the system once that it was operational.

In the second event, a panel at ECDL 2008, it was discussed the relationships between the Web and both the traditional and the digital libraries. To stimulate discussion, the view of one camp was claimed to be that since “all” the information was available on-line, the use of smart search engines and clever software tools would allow the Web to provide all the information (and the services) needed by an information seeker. The view of the other camp was that the value of information was not just in its sheer quantity, but was rather in the organization and the quality of the information made available, and that could never be accomplished by “programs”. Some years later, with the continuous increase of the information available on the Web and the advances in search engine technologies, an even more radical question could be raised, questioning the need of libraries at all, whether digital or not. More and more it appears that when there is an information need, everybody (including scholars) is first “googling” on the Web to find the desired information, and it is not known how many information seekers will continue by accessing also some (digital) libraries in order to satisfy their information need. During these years however, digital library technologies have supported the transition of libraries from traditional to digital, and those technologies are today mature enough to support not only the availability of the library content online, but also the provision of advanced services for library users.

At the end of the panel the position that gathered most of the consensus was the one supported (not surprisingly) by Google, The main mission of a web search engine should be to provide access to the ”world’s information”, and make it universally accessible and useful, whereas the main mission of a digital library should be to organize the information needed and used by one or more specific user communities and make it easily accessible and useful to those communities. The difference in mission implies therefore a difference in scale (the web is measured in billions of pages, a digital library is measured, at best, in millions of documents), a difference in coverage (as broad as possible in the web, as deep as possible in the library) and a difference in services, i.e. how to add value to the content of the library (precision and general services in the web, completeness and specific services for a user community in the library). The web and the digital library have therefore similar and complementing

missions, and they should take advantage of each other, and focus on the delivery of useful and relevant (web) services to their user communities.

What emerged from these first two events was the identification of three main profiles at the operational level of a library. Two of them, namely the digital librarian and the system librarian, have been mentioned before , while the third one, that could be called the "end-user librarian", is a profile with a deep knowledge of the information needs and applications of the selected user community. The end-user librarian should be able to provide input to the digital librarian on one side and to assist the library users on the other, by providing reference services (possibly using web search engines) and assistance in the use of the new functionality (possibly) made available by the digital library, such as annotations and collaboratories.

5 Research in Digital Library

In 2010 the International Master DILL (DIgital Libraries Learning) and the European project DL.org organized together a one-day seminar ("Research & Education in Digital Libraries"), as a forum for discussion between the research communities participating in the DL.org activities and the communities of Digital Library education in Europe, with the aim of starting a dialogue about research and education in digital library and to explore ways for a closer cooperation between those communities. DILL is a two-year international master program (which was funded until 2011 by the European Union under the Erasmus Mundus program) that is bringing forward the idea of interdisciplinary education in Digital Libraries by providing to its students courses which span some of the different aspects underlying digital libraries. DL.org (now ended) was a project funded by the European Union under the 7th Framework Program to bring forward a research program focussed on interoperability in digital libraries, which means that research should consider not only the technical dimension, but also other dimensions that might be affected by interoperability issues (e.g. policy, quality, user profiles, legal aspects).

Among the main accomplishments of the DL.org project there is the completion of a conceptual model for Digital Libraries [3] (initiated by the DELOS Network of Excellence), which includes the three roles (profiles) of library professionals mentioned above, and which has been widely used in DILL and in other courses and Summer Schools addressing the educational needs of library professionals, to establish a common view of the entities and the concepts underlying the "Digital Library Universe".

The stated objectives of the seminar (only partially attained) were:

- Start discussing how to implement a European scale mechanism for exchanging, sharing and integrating research results into education in digital libraries
- Start defining research topics suitable for PhD students to ease the integration of research done in European projects and research done in Universities
- Discuss how the interoperability research results of DL.org can be transferred to education in digital libraries

The discussions prompted by a number of interesting presentations brought into evidence a wide range of issues, going from the need to transfer research outcomes into learning material, to the need for DL professionals to have hands-on experience with IT tools and services, to the need to work towards stronger theoretical foundations for digital libraries. A practical result was the identification of a few research topics for DILL Master Thesis and the opportunity for internships for the DILL students attending the workshop.

In 2011 DL.org and DILL continued their cooperation organizing a workshop ("Linking Research and Education in Digital Libraries") in connection with the conference TPDL 2011. The aim of the workshop was to bring forward the discussions already started at the previous events, namely how to better exploit the results of research for education in Digital Libraries, or more generally, for education to "information workers". As briefly mentioned at the beginning, all professionals working in the so-called "memory institutions" (libraries, archives and museums) are increasingly facing the need to reconsider their educational needs in order to maintain the traditional leadership in the cycle of knowledge creation, distribution and preservation. The increased availability of digital information made possible by the Web is blurring the boundaries between those institutions and is transforming the respective professionals in a more general role of "information workers". The main thread of discussion was a critical review of the roles of the information professionals, considering not only the impact brought by the advances in the technical dimension, but considering also other dimensions such as policy, quality, user profiles, legal aspects, etc.

A number of interesting topics were presented and discussed during the workshop, such as the need for a theoretical foundation in order to transform the "librarian profession" into a "librarianship discipline"; the possibility of using a "conformance checklist" to assess the conformance of a digital library with the conceptual model proposed by DL.org, showing how the checklist could provide the basis for defining a set of topics needed in digital library education. A discussion about the skills needed by a professional in order to evaluate a digital library, focusing more on the organizational and interpersonal skills rather than the technical ones, was useful in highlighting a (different) set of topics needed in digital library education. Several examples advocated the early involvement of students into research projects requiring skills both in Library and Information Science and in Computer Science. An interesting perspective introduced the notion that a change in terminology, when going from the library world to the world of the Web and Linked Open Data (e.g. from "catalogue" to "graph", from "document" to "aggregation"), actually implies a complete re-thinking of the meaning of all those terms, and therefore also a re-thinking of their educational aspects.

A concluding panel provided additional views and experiences in education in digital libraries. At the conclusion of the panel and the of workshop there was a general agreement that information professionals, given the increased use of Web technologies for knowledge dissemination and for collaboration, definitely need an increased education in the usage (and development) of interactive tools and services to

facilitate their activities as information professionals. It was unclear (and it was left open) how and where to draw the line between increased education in Computer Science in general, and increased education in the usage of advanced applications and tools available for memory institutions.

In 2012 DL.org and DILL organized together another event in this series, namely a panel ("Can Research help Education in Digital Libraries ?") in connection with the conference LIDA 2012. As in the other events, the main aim of the panel was to explore how the research activities and the educational activities can interact together at an earlier stage, in order to benefit each other from a better knowledge of the respective needs and objectives. The panellists were chosen so as to represent both sides of the matter. The introduction to the panel presented the following considerations, as a way to start a debate both among the panellists and with the audience.

It is becoming more and more clear that the pace of advancement of the technologies underlying and supporting Digital Libraries and the services that they provide is not matched by similar advancements in the educational curricula leading to "Digital Librarianship". Over the last 15 years Digital Libraries, or more generally the "Memory Institutions", have seen a significant level of research in many of the fields that in one way or another are related to the production, description, collection, preservation, retrieval and usage of digital information. In many cases the outcome of those research activities has resulted in tools and technologies (e.g. interoperability of data at the semantic level, natural language processing, automatic analysis and classification of texts, building of multimedia collections) which allow a more effective way of providing the traditional services of the memory institutions.

In parallel with those developments, the educational curricula of librarians, archivists and museum curators have been (slowly) updated to reflect the changes in the professional environment, but those changes in the curricula often appear to be dictated more by the need to "run after" the technology, rather than a deep re-thinking of the educational needs of the memory professionals, resulting just in the increase of the "technology component" of the curricula.

The presentations and the discussions during the panel somehow confirmed those initial considerations, especially when looking at some of the emerging areas of interest for the publication, access and re-use of scientific material. For example, there is an increasing need to publish and make accessible experimental data (e.g. data banks, data sets of results, methodologies and workflows), which implies for a library the ability to manage new and different types of content; there is an increasing need to curate, collect, aggregate and make available data coming from many different sources, which implies for a library the ability to manage repository registration and validation, policy definitions, representation of ontologies and mappings, etc.; there is the need to provide access to different "views" of a digital library, which implies for a library the ability to provide "virtual digital libraries" on demand. Very few of the participants in the panel and the audience seemed to have "standard" curricula covering those emerging topics.

6 Conclusions

Coleman [7] writes that often the starting place for designing an interdisciplinary course involves an eight-step process to interdisciplinary course and curriculum planning:

1. Assemble an interdisciplinary team;
2. Select a topic;
3. Identify disciplines from which the course needs to draw;
4. Develop the subtext for the course (subtext is the abstract issue or issues which form the substantive topic of the course);
5. Structure the course by identifying the conceptual glue that holds it together, keeping in mind not only what is taught but to whom;
6. Select the readings;
7. Design the assignments;
8. Prepare the syllabus. The syllabus must specify what disciplines are included and why.

Through the series of events described in the paper, the interdisciplinary collaboration between Library and Information Science and Computer Science has been able to achieve a preliminary understanding of steps 2 and 3 and (to some extent) 4 of this process, focusing on the Digital Library domain. The events were useful for identifying the state and characteristics of education and theoretical research in Digital Library and confirmed the understanding that both theory building and theory use in education are intertwined, in order to construct a cohesive body of knowledge in the field. The results confirm that the degree of interdisciplinarity within Digital Library has increased and is growing. Further research is needed to evaluate this and other strategies based on the recognition of a wider range of channels for communication of research to practice and education. The events showed a tendency to converge into a few subfields, such as digital curation, information seeking and Digital Library use, information retrieval in the Web. However, the declining share of theoretical developments are showing to Library and Information Science researchers the urgency and the importance of continuous and creative research in that field, in collaboration with Computer Science researchers.

References

1. Tammaro, A.M., Casarosa, V., Borgman, C., Connaway, L.S., Castelli, D., Radford, M. (panellists): Can Research help Education in Digital Libraries?, http://ozk.unizd.hr/lida/program
2. Bruce, B.C., Kaptizke, C.: Libr@ries: Changing information space and practice. Lawrence Erlbaum, Hillsdale (2006) ISBN 0-8058-5481-9

3. Candela, L., Castelli, D., Pagano, P., Thanos, C., Ioannidis, Y.: Setting the foundations of digital libraries: the DELOS
4. gital-librariesmanifesto. D-Lib Magazine 13(3/4) (March/April 2007), http://www.dlib.org/dlib/march07/castelli/03castelli.html
5. Carr, N.: The Big Switch: Rewiring the World, from Edison to Google. W. W. Norton, New York (2008)
6. Casarosa, V., Castelli, D., Tammaro, A.M.: Report on the Workshop "Linking Research and Education in Digital Libraries". D-Lib Magazine 17(11/12) (November/December 2011), http://www.dlorg.eu/index.php/dl-org-events/tpdl-workshop-linking-research-education-in-dlsn-in-dls, doi:10.1045/november2011-casarosa
7. Casarosa, V., Cousins, J., Tammaro, A.M., Ioannidis, Y.: The Web Versus Digital Libraries: Time to Revisit This Once Hot Topic. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 383–384. Springer, Heidelberg (2008), http://dx.doi.org/10.1007/978-3-540-87599-4_39, http://www.ecdl2008.org/panels/
8. Colemann, A.: Interdisciplinarity: The Road Ahead for Digital Library Education. D-LibMagazine (8) (July-August 2002), http://www.dlib.org/dlib/july02/coleman/07coleman.html
9. Computer Science Curricula 2013 (CS2013). ACM/IEEE-CS Joint Task Force (2013), http://ai.stanford.edu/users/sahami/CS2013/ironman-draft/cs2013-ironman-v0.8.pdf
10. DL.org and DILL, Workshop Research and Education for Digital Libraries, Parma (November 2010), http://www.dlorg.eu/index.php/dl-org-events/research-education-in-di
11. IEEE Technical Committee on Digital Libraries (1997), Position Statement, http://www.ieee-tcdl.org/mediawiki/TCDL/index.php/Position_Statement
12. Information Technologies Profiles and Curricula for Libraries, Parma, Italy - Sala Conferenze Oratorio Novo, Vicolo S. Maria 5, October 13-14 (2005), http://www.unipr.it/arpa/benicult/biblio/master/131005.htm
13. Koenig, M.D., Hildreth, C.: The End of the Standalone "Library School". Library Journal (June 15, 2002), http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA220870
14. IFLA. Guidelines for Professional Library/Information Educational Programs (2012), http://www.ifla.org/VII/s23/bulletin/guidelines.htm
15. Miller, W., Pellen, R.M.: Libraries and Google Haworth Information Press (2005)
16. Myburgh, S., Tammaro, A.M.: Education for Digital Librarians: Some European Observations. In: Spink, A., Heinström, J. (eds.) Library and Information Science Trends and Research: Europe (Library and Information Science, vol. 6), pp. 217–245. Emerald Group Publishing Limited (2012)
17. Pomerantz, J., Abbas, J., Mostafa, J.: Teaching Digital Library Concepts Using Digital Library Applications. International Journal on Digital Libraries 10(1), 1–13 (2009)
18. Pomerantz, J., Oh, S., Yang, S., Fox, E.A., Wildemuth, B.: The Core: Digital Library Education in Library and Information Science Programs. D-Lib Magazine 12(11) (2006)
19. Spink, A., Cool, C.: Education for Digital Libraries. D-Lib Magazine 5(5) (May 1999), http://www.dlib.org/dlib/may99/05spink.html

20. Tammaro, A.M.: IT profiles and curricula for digital libraries in Europe. LIDA (2006), http://dspace-unipr.cilea.it/bitstream/1889/1185/1/Tammaro_LIDA_2006.pdf
21. Tammaro, A.M.: A curriculum for digital librarians: a reflection on the European debate. New Library World 108(5/6), 229–246 (2007)
22. Tennant, R.: Digital Libraries: The Digital Librarian Shortage. Library Journal (March 15, 2002), http://libraryjournal.reviewsnews.com/index.asp?layout=article&articleid=CA199859&display=Digital+LibrariesNews&industry=Digital+Libraries&verticalid=151

Author Index

Agosti, Maristella 147
Aloia, Nicola 53
Artini, Michele 77
Atzori, Claudio 77

Bardi, Alessia 77
Bertazzo, Matteo 59
Biba, Marenglen 17

Casarosa, Vittore 81, 187
Castelli, Donatella 187
Catarci, Tiziana 7
Ceci, Michelangelo 29
Concordia, Cesare 53
Conlan, Owen 41

Di Iorio, Angela 59

Esposito, Alfredo 70
Esposito, Floriana 17

Felle, Antonio E. 29
Ferilli, Stefano 17, 93
Ferrara, Felice 105
Ferro, Nicola 41, 130
Fumarola, Fabio 29

Grieco, Domenico 17
Guercio, Maria 7, 59

Hampson, Cormac 41

La Bruzzo, Sandro 77
Leuzzi, Fabio 93

Magarotto, Adriana 165
Malerba, Donato 29
Manfioletti, Marta 147
Manghi, Paolo 77
Manoni, Paola 1
Martinoli, Adriana 70
Meghini, Carlo 53, 81
Mikulicic, Marko 77
Munnelly, Gary 41

Orio, Nicola 147
Ortolani, Silvia 59

Peroni, Silvio 118
Pio, Gianvito 29
Ponchia, Chiara 147

Quaquarelli, Maura 165

Raieli, Roberto 171
Rotella, Fulvio 93

Sadeh, Tamar 153
Santucci, Giuseppe 7
Schaerf, Marco 59
Silvello, Gianmaria 130, 147

Tammaro, Anna Maria 136, 187
Tasso, Carlo 105
Tomasi, Francesca 7, 118
Trupiano, Luca 53

Vallania, Mattia 165
Vitali, Fabio 118

Zengenene, Dydimus 81
Zingoni, Jacopo 118
Zoppi, Franco 77